# Electrostatic Models for Implicit Solvation and Applications to Protein Folding and Aggregation

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Urs Edgar Haberthür

von

Winterthur ZH

Begutachtet von

Prof. Dr. Amedeo Caflisch

Zürich 2004

Die vorliegende Arbeit wurde von der Mathematisch-naturwissenschaftlichen Fakultät der Universität Zürich auf Antrag von Prof. A. Caflisch und Prof. H. R. Bosshard als Dissertation angenommen.

*Meinen Eltern gewidmet*

*What you see is what you see.*

Frank Stella

# Publications

## Articles in International Peer-Reviewed Scientific Journals

**U. Haberthür, A. Caflisch**
FACTS: Fast Analytical Continuum Treatment of Solvation
Submitted

**U. Haberthür, N. Majeux, P. Werner, A. Caflisch**
Efficient Evaluation of the Effective Dielectric Function of a Macromolecule in Aqueous Solution
J. Comput. Chem. 2003, 24, 1936-1949

**A. Cavalli, U. Haberthür, E. Paci, A. Caflisch**
Fast Protein Folding on Downhill Energy Landscape
Prot. Sci. 2003, 12, 1801-1803

**J. Gsponer, U. Haberthür, A. Caflisch**
The Role of Side-Chain Interactions in the Early Steps of Aggregation: Molecular Dynamics Simulations of an Amyloid-Forming Peptide from the Yeast Prion Sup35
Proc. Natl. Acad. Sci. USA 2003, 100, 5154-5159

## Molecular Dynamics Simulation Code in CHARMM

**U. Haberthür**
The FACTS Implicit Solvation Model
Submitted

**U. Haberthür**
The SASA Implicit Solvation Model
Available in CHARMM since 2001, version c28b1

# Summary

Understanding how a protein folds from a linear chain to its native state remains a major challenge in biology. Substantial progress in experimental and theoretical studies has been achieved over the last years. As a result a qualitative view of the protein folding process could be established. Yet, quantitative models that elucidate the molecular bases of folding are still missing. Moreover, a significant number of pathologies is related to protein misfolding which motivates additional research. The present thesis focuses on two issues. The first one is the development and implementation of computational methods that facilitate the study of structural changes in biological molecules on a computer. The second one is the application of these methods to current problems of folding and aggregation.

Sampling efficiency is a key issue of molecular dynamics (MD) simulations of biological compounds on a computer. The explicit numerical integration of Newton's equations of motion permits a maximal time step of only 1-2 fs. This typically limits the length of all-atom MD simulations to nanoseconds. In contrast, many biologically relevant processes like protein folding take between microseconds and minutes. Furthermore, to obtain statistically relevant results several MD simulations of the same process have to be performed.

A major reason for the unsatisfactory timescale of MD simulations is solvation. Solvation plays a crucial role in protein structure and function. Hence, it can not be neglected in MD simulations although the detailed motion of the solvation molecules is only rarely of interest. It is therefore particularly frustrating that about 90% of the computation time has to be spent on the exact calculation of the solvent molecules. Implicit solvation models provide a powerful way to avoid this problem. They eliminate the solvent degrees of freedom with the goal to be thermodynamically equivalent to the explicit treatment. Implicit models incorporate the solvation effects into a mean solvation energy term that is added to the vacuo potential energy function of the solute.

The first part of this thesis focuses on the development and implementation of implicit solvation models. Continuum electrostatics is applied to approximate the mean solvation term. Two models, AEI and its successor FACTS, are developed from scratch, parameterized, and implemented in CHARMM. They represent a combination of speed and accuracy on a level that has not been reported before. The AEI and FACTS models are only about three to four times slower compared to MD simulations without solvation and compete in accuracy with schemes that are between twenty to forty times slower than simulations in vacuo. The SASA model is another fast implicit solvation model. It is based on the solvent accessible surface area and

is less rigorously founded on continuum electrostatics than the AEI and FACTS approaches. The SASA model has been implemented in the official CHARMM release as part of this thesis and is used in several studies.

The second part of the thesis consists of applications of implicit solvation models to simulations of protein folding and aggregation. In the first study the reversible folding of a three-stranded antiparallel $\beta$-sheet peptide is reported. During a total simulation time of 12.6 $\mu$s at the melting temperature, 72 folding events are observed. It is demonstrated that the unfolded state ensemble contains many more conformers than those sampled during a single folding event. This confirms previous findings in lattice models that fast folding corresponds to a downhill process on a funnel-like free energy surface. The second study investigates the role of side-chain interactions in the early steps of the amyloid fibril forming process. Aggregation MD simulations of prion-like peptides of the yeast protein Sup35 are performed. In agreement with experimental data they correctly generate the in-register parallel packing of $\beta$-strands. Backbone interactions favor the in-register antiparallel packing. In contrast, hydrogen bond interactions between side-chains and stacking of aromatic residues favor the in-register parallel assembly. Overall, the parallel side-chain interactions are slightly dominant over the antiparallel backbone interactions and determine the statistically most relevant configuration. The roughness of the effective and free energy surfaces is shown and it is demonstrated that the preferred pathway to the parallel aggregate does not correspond to a purely downhill profile on the effective energy surface. Additionally, the simulations confirm a strong sequence dependence of the aggregation kinetics.

# Zusammenfassung

Das Verständnis für den Prozess, wie sich ein Protein von einer linearen Kette zu seinem nativen Zustand faltet, stellt nach wie vor eine grosse Herausforderung für die Biologie dar. In den letzten Jahren konnten grosse Fortschritte sowohl bei den experimentellen Techniken als auch bei den theoretischen Modellen erreicht werden. Als Folge etablierte sich eine qualitative Sicht des Proteinfaltungsprozesses. Quantitative Modelle, welche die molekulare Ebene des Faltungsprozesses beschreiben, fehlen aber noch immer. Zudem stimuliert die bedeutende Anzahl an Pathologien, die mit einem fehlerhaften Faltungsprozess in Zusammenhang gebracht werden können, das Interesse an der Proteinfaltung zusätzlich. Die vorliegende Arbeit umfasst zwei Schwerpunkte. Einerseits werden mathematische Modelle, die helfen, strukturelle Veränderungen biologischer Moleküle auf dem Computer zu simulieren, entwickelt und in Computerprogrammen implementiert. Andererseits werden diese Modelle auf aktuelle Fragestellungen der Faltung und Aggregation angewandt.

Ein genügend grosses Sampling zu erreichen ist eine der Hauptschwierigkeiten von Simulationen der Molekül̈dynamik (MD) biologischer Verbindungen auf dem Computer. Die explizite numerische Integration der Newtonschen Bewegungsgleichungen erlaubt einen maximalen Zeitschritt von nur 1-2 fs. Dies limitiert die klassische MD-Simulation, wo alle Atome explizit abgebildet werden, auf eine Länge im Bereich von Nanosekunden. Dem steht die Tatsache gegenüber, dass viele biologisch relevante Prozesse wie die Proteinfaltung Mikrosekunden bis Minuten benötigen. Erschwerend kommt hinzu, dass, um statistisch signifikante Resultate zu erhalten, mehrere MD-Simulationen des gleichen Systems erforderlich sind.

Einer der Hauptgründe für die unbefriedigende Länge von MD-Simulationen ist das Lösungsmittel. Die Solvatation ist entscheidend für Struktur und Funktion eines Proteins. Deshalb darf sie in MD-Simulationen nicht vernachlässigt werden, obwohl die genaue Bewegung der Lösungsmittelmoleküle nur sehr selten von Interesse ist. Es ist deshalb besonders frustrierend, dass etwa 90% der Rechenzeit für die detailgetreue Abbildung der Solvatationsmoleküle verwendet werden muss. Implizite Solvatationsmodelle bieten eine hervorragende Möglichkeit, dieses Problem zu vermeiden. Sie eliminieren die Freiheitsgrade des Lösungsmittels mit dem Ziel, eine zur expliziten Behandlung thermodynamisch äquivalente Beschreibung zu liefern. Ein implizites Solvatationsmodell vereinigt die Effekte des Lösungsmittels in einem gemittelten Solvatationsenergieterm, der zur potentiellen Energie des Proteins im Vakuum addiert wird.

Der erste Teil dieser Dissertation befasst sich mit der Entwicklung und Implementierung von impliziten Solvatationsmodellen. Die Theorie der Kontinuumselektro-

statik wird benutzt, um den gemittelten Solvatationsenergieterm approximativ zu berechnen. Zwei Modelle, AEI und dessen Nachfolger FACTS, werden von Grund auf entwickelt, parametrisiert und in CHARMM implementiert. Diese erreichen eine Kombination von Geschwindigkeit und Genauigkeit auf einer Stufe, wie sie bisher noch nicht publiziert wurde. Die AEI- und FACTS-Modelle sind nur etwa drei- bis viermal langsamer als MD-Simulationen ohne Solvatation, konkurrieren in Genauigkeit aber mit Modellen, die zwanzig- bis vierzigmal langsamer sind als MD-Simulationen ohne Lösungsmittel. Ein weiteres schnelles Solvatationsmodell ist das SASA-Modell. Es basiert auf der Berechnung desjenigen Teils der Oberfläche eines Proteins, die dem Lösungsmittel zugänglich ist, und baut insgesamt weniger streng als die AEI- und FACTS-Modelle auf der Theorie der Kontinuumselektrostatik auf. Das SASA-Modell wurde als Teil dieser Dissertation in die offizielle Ausgabe von CHARMM implementiert und kommt in mehreren Studien zur Anwendung.

Der zweite Teil der Dissertation befasst sich mit Anwendungen impliziter Solvatationsmodelle auf die Proteinfaltung und die Aggregation. In der ersten Studie wird der reversible Faltungsprozess eines dreisträngigen $\beta$-Faltblattes simuliert. Während der gesamten Simulationszeit von 12.6 $\mu$s, durchgeführt bei der Schmelztemperatur, werden 72 Faltungen beobachtet. Es wird gezeigt, dass das Ensemble des entfalteten Peptids bedeutend mehr Cluster enthält als die, welche während eines einzelnen Faltungsprozesses tatsächlich besucht werden. Dies bestätigt Resultate von Gittermodellen, die besagen, dass die schnelle Faltung einem Prozess entspricht, der auf einer einem Trichter ähnlichen freien Energiefläche bergab verläuft. Die zweite Studie untersucht die Rolle der Interaktionen zwischen Seitenketten im Frühstadium der Amyloidbildung. Aggregationssimulationen von prionähnlichen Peptiden des Yeast-Protein Sup35 liefern Resultate, die mit experimentellen Daten übereinstimmen. Sie erzeugen die korrekte parallele Struktur der $\beta$-Stränge. Zwar bevorzugen die Wechselwirkungsenergien zwischen den Rückgraten eine antiparallele Anordnung. Die Wechselwirkungen zwischen den Seitenketten infolge Wasserstoffbrücken und der Aneinanderlagerung aromatischer Residuen favorisieren jedoch die parallele Struktur. Insgesamt sind die Wechselwirkungsenergien in der parallelen Anordnung leicht günstiger als in der antiparallelen, was die statistisch dominante Konfiguration bestimmt. Des weiteren wird die Zerklüftung der effektiven und freien Energieflächen veranschaulicht, und es wird gezeigt, dass der bevorzugte Weg zu einem parallelen Aggregat nicht einem Profil auf der effektiven Energiefläche entspricht, das ausschliesslich bergab verläuft. Zudem bestätigen die Simulationen eine starke Abhängigkeit der Aggregationskinetik von der Sequenz.

# Contents

# Part I

# Overview

# Chapter 1

# Simulation Science: Between Theory and Experiment

## 1.1 The nature of science

The very nature of science is to understand how the world works. It is the quest for the fundamental laws and their interplay in order to conceive even the most complex processes. From the smallest dimensions of elementary particles to the astronomical extensions of the universe the human mind seeks for the causal connections. The cornerstone of modern science as we experience it today was laid by the so called academies. They were founded all over Europe during the course of the 17th and 18th century. The academies signalized the advent of the new spirit of the age of enlightenment which had its seeds in the polemic fight against religion. Since that time modern science has evolved dramatically and found its present peak level in the 20th century. The impact of the scientific discoveries on nearly all areas where humans are involved can not be overemphasized. Einstein's special theory of relativity, published 1905, Heisenberg's uncertainty principle, published 1925, parity violation, discovered 1959, to name just a few, thrilled the established world view. They had fundamental and irreversible implications not only for research, development, and technology, but also for philosophy. While the 20th century is sometimes called the age of physics among scientists the next era is expected to be the era of life science and nanotechnology.

The building blocks of science are models. A model is a simplified and rational description of a part of reality. It is designed to provide insights and answers at a given abstraction level, and the more detailed the model, the more detailed the output. On the one end there are the solely qualitative models. Their aim is to give a basic idea of a certain process. On the other end models are so sophisticated

that their predictions and the experimental results differ only within the measuring error. Building a model is a well recognized way of understanding the world and to predict what will happen under given initial conditions.

## 1.2 The classical approach

The key in the classical scientific approach is the experiment in a physical laboratory ("cut and try"). Depending on the phenomenon under investigation and the knowledge already available an experiment is designed and conducted. The database extracted from the results is the basis for constructing a new model or refining and validating an already existing one. In the early days of research single individuals planned and performed experiments, draw conclusions, and built models. One example is the monumental research work of Coulomb that began to be published in 1785. With increasing complexity of both experiments and models researchers began to split up in experimentalists and theorists. Experimentalists are devoted to design and conduct experiments whereas theorists are concerned with creating and refining models. Clearly, experimentalists must have a profound knowledge of the underlying theory in order to plan the most fruitful experiments. On the other hand, theorists should have a great deal of understanding about the experiments in order to interpret data correctly and make suggestions for further measurements. Some of the most outstanding examples of how demanding in terms of human resource, energy, time, and money research has become nowadays are the experiments at CERN, the world's largest particle physics laboratory. Thousands of scientists work together, founded by a conglomerate of many countries, to do research on elementary particles. Such tremendous effort sometimes has unexpected side effects. The Internet, for instance, is an heritage of a network originally designed and introduced at CERN.

## 1.3 The modern approach

The rapid and simultaneous progress in computer technology and software in the second half of the 20th century added a third category to the experimentalist and theorist: the simulation scientist. Simulation science is interdisciplinary and can be defined as the intersection of numerical mathematics, computer science, and modeling [1]. The key to the enormous power of simulation science is to recognize that due to modern computer technology mathematical models can be solved numerically and no longer need to be solved analytically. This opens the door to a huge number

of applications that were out of range until recently. Most mathematical models can only partially or not at all be solved analytically. The more precisely they describe the processes measured in experiments, the more complex they grow and the differential equations are no longer integrable. This is where simulation science takes its place between experiment and theory. On the one hand it imitates an experiment, and on the other hand it can only do so if a mathematical description of the process under investigation is already available. Simulation science complements or replaces the material experiment in a physical laboratory by a numerical experiment on a computer, sometimes called the virtual laboratory. This marks the step from the classical scientific approach of "cut and try" to the modern approach of "simulate and analyze". Since the numerical experiment is founded directly on the mathematical model simulation science is closer related to theory than an experiment. The modeling phase is as important to the simulation scientist as the numerical experiment itself.

While the experimentalist deals with matter and the theorist with paper and pencil, the simulation scientist deals with bits. This requires an adaption of the classical research procedure. A short description of the various steps in the simulation approach is therefor appropriate. The first task is to devise a mathematically and physically consistent model and a clear formulation of all assumptions. This includes, for instance, mathematical rigor and compliance with conservation laws. The first task can be called modeling and requires knowledge in mathematics, physics, and in the case of protein folding also biochemistry. The next step is to develop a suitable algorithmic procedure. Questions like how to discretize a continuum model or what kind of integrator fits the task at hand best have to be answered. This kind of problems is assigned to the realm of numerical mathematics. The third task is to convert the numerical algorithms to software that computes efficiently. How efficient the calculations can be done translates directly to how sophisticated a model can be simulated. The third task requires knowledge of an appropriate computer language, algorithms and data structures, debugging, and the like. Clearly, these problems belong to the world of computer science. The fourth and final step is to analyze and assess the accuracy of the results, to visualize, and eventually publish them.

# Chapter 2

# Protein Folding

## 2.1 Introduction

Proteins (Greek *proteios*, of first importance) are macromolecules that play a central role in nearly every biological process. All proteins are synthesized from the same twenty amino acids as linear chains whose sequences are encoded in the DNA. In order to fulfill its function a protein folds under physiological conditions to a unique and stable three-dimensional structure within a biologically relevant time. Understanding how a protein finds its way from a linear chain to its native state is called the protein folding problem and is one of the most challenging issues in biology. The folding process is of great complexity since it is the transformation from a highly disordered to a highly ordered state. Thousands of non-covalent interactions need to be established. The experiments by Anfinsen [2] demonstrated that the native state of a protein is determined solely by its amino acid sequence and not by other environmental factors that would be hard to characterize. This key discovery detached the protein folding problem from its complex cellular environment and made it amenable to theoretical studies.

A large amount of experimental and theoretical work on protein folding and related subject has been done in the past fifty years (for a review see [3] and [4]). Substantial progress in both areas has been made in the last two decades [5]. On the experimental side new methods were established that allow to extract data from processes on the nanosecond timescale. On the theoretical side models that are simplified such as to be tractable on a computer but still complex enough to be meaningful gave basic insights in the mechanism of protein folding. The experimental and theoretical work resulted in a consistent and commonly agreed on view of how proteins fold. A short survey of the more recent experimental and theoretical methods that address the protein folding process is given in the following.

Afterwards the current qualitative picture of protein folding is described [6].

## 2.2 Experimental Studies

Both time and space resolution achieved in experiments have been dramatically increased in recent years. Fast folding techniques allow to probe the milli-, micro-, and even nanosecond timescale. Two ingredients are necessary for fast folding experiments: a rapid initiation of the folding process, followed by a rapid probing. The key concept of a rapid initiation is to introduce a sudden change in the solvent environment of the protein that shifts the equilibrium constant. As a consequence the concentration relaxes either toward higher concentration of folded proteins (folding experiments) or toward higher concentration of unfolded proteins (unfolding experiments).

The most important fast initiation techniques are mixing, photochemical triggering, and temperature jump. Although all these techniques were already developed in the 1970s, applications to folding had to wait until the 1980s (with the exception of homopolymers). Mixing was the earliest relaxation technique applied to fast protein folding. The protein solution is combined with a buffer and thus the equilibrium constant is shifted. Conventional stopped-flow mixers give a time resolution of several milliseconds. The more recent continuous mixers go below 100 $\mu$s. Stopped-flow mixing experiments were among the first ones to reveal fast two-state folding without the occurrence of intermediates [7]. In photochemical triggering a laser pulse is applied that induces a chemical change which in turn shifts the equilibrium constant in a time range from pico- up to microseconds. It has been used, among other things, to study diffusional contact times of backbone segments as a function of length [8]. Temperature jump techniques rely on a sudden change in temperature triggered by a laser pulse. They have been applied to study protein folding, unfolding, and secondary structure formation [9, 10].

The most important folding probe techniques include amide exchange, NMR lineshapes, circular dichroism, direct absorption, fluorescence, and resonance Raman. Amide exchange allows to identify folded regions and regions that are either completely unfolded or extensively fluctuating between secondary structure and random coil [11]. The time resolution can go down to 5 ms. NMR lineshapes are used to study two-state folding. These experiments are actually conducted under equilibrium condition and are not combined with rapid initiation. NMR lineshapes unambiguously establish the two-state nature of the folding equilibrium. The time resolution is between 100 $\mu$s and 50 ms. These techniques facilitated one of the

first direct measurements of a sub-millisecond two-state folding rate [12]. Circular dichroism in the far UV (190-230 nm) can monitor the formation of secondary structure elements. In the near UV (250-300 nm) it delivers information on the packing of aromatic side-chains [13]. The time resolution goes down to the micro- and nanosecond timescale. Direct absorption allows to examine both secondary structure formation [14] and loss [15] on a millisecond timescale. Fluorescence also probes the millisecond timescale and delivers information on global kinetics, solvent exposure, motional anisotropy, formation of specific tertiary contacts, and distance correlation functions. The method has been used to measure collapse times of unfolded proteins [16, 17] and to give accurate three-dimensional information during folding by distance measurements of pairs [18]. Such data on a millisecond timescale can provide valuable information on the average three-dimensional history of protein folding. Resonance Raman probes the millisecond and sub-millisecond timescale and can be used to determine secondary structure content.

Obtaining structural information on the rate limiting step or transition state, in particular for simple two-state uni-molecular folders, is another important link in the chain to understand the mechanism of protein folding [19]. The protein engineering method, pioneered by Fersht and coworkers, allows to obtain an accurate description of the transition state ensemble [20]. The idea is to first apply a point mutation to a single residue, followed by measuring free energy shifts of the native and the transition state relative to the unfolded state. This allows to probe whether the neighborhood of the mutated residue in the transition state is similar to the neighborhood in the native or the unfolded state. More precisely, the $\phi$ value of a residue is defined by

$$\phi = \frac{\Delta G^w_{T-U} - \Delta G^m_{T-U}}{\Delta G^w_{F-U} - \Delta G^m_{F-U}} = \frac{\Delta\Delta G_{T-U}}{\Delta\Delta G_{F-U}}$$

where $\Delta G^w_{T-U}$ is the free energy difference between the transition and unfolded state of the wild type, and $\Delta G^w_{F-U}$ is the free energy difference between the folded and unfolded state of the wild type. $\Delta G^m_{T-U}$ and $\Delta G^m_{F-U}$ are the analogous quantities for the mutant. If the neighborhood of the mutated residue in the transition state is similar to its neighborhood in the native state, both the transition and native state are shifted by a similar amount of free energy relative to the unfolded state and $\phi$ is close to 1. If the neighborhood of the mutated residue in the transition state is similar to its neighborhood in the unfolded state, then only the native but not the transition transition state is shifted relative to the unfolded state and $\phi$ is close to 0. Values larger than 1 or lower than 0 are generally attributed to non-native

contacts or to the existence of several parallel rate limiting steps. The transition state ensembles of chymotrypsin inhibitor 2 (CI2) and different SH3 domains were mapped out in detail with the protein engineering method [21, 22, 23]. Most of the $\phi$-values are fractional which suggests that the transition state is essentially an extended version of the native state. It was concluded that folding involves to first bring together a nucleus of key residues, followed by condensation of the rest of the structure around the core. Accordingly the term nucleation-condensation was introduced for this kind of folders. Furthermore, it was shown that the transition state ensembles of the src and $\alpha$-spectrin SH3 domains are very similar although the sequence identity is only around 35% [22, 23]. This led to the conclusion that the native state per se implies certain mechanisms for the folding process [24].

## 2.3 Theoretical Studies

Experiments deliver structural, thermodynamic, and kinetic information on specific proteins. Theoretical models try to capture a more global view of the protein folding process. Some of the most valuable insights into how a protein folds has been gained by lattice models [25, 26, 27]. The basic idea of a lattice model is to reduce the size of the conformational space available to a real protein such that all thermodynamic quantities can be calculated on a computer within a sensible time but that an exhaustive search through all microstates is still not possible. In a lattice model each amino acid is represented by one or two beads on either a discrete grid ("on lattice") or in a continuum space ("off lattice"). The potential energy of the system is a simple function of the contacts between the beads. The simulations are carried out by Monte Carlo steps in order to provide a Boltzmann distributed sampling in accordance with the canonical ensemble. For such models it is possible to clearly formulate the necessary and sufficient condition for fast folding and thermodynamic stability of the native state. This single condition is that the native state be a pronounced global minimum on the potential energy surface. Other suggested mechanisms for folding could not be confirmed. The potential energy includes all contributions to the free energy energy except the conformational entropy. Since the simplified potential energy function of lattice models neglects non-native interactions, their potential and free energy landscapes are very smooth. In real proteins non-native interactions introduce roughness into these energy landscapes.

This roughness can be reproduced in a more detailed picture of the thermodynamics and kinetics of protein folding that is given by molecular dynamics (MD) simulations. The underlying model is an all-atom representation of the protein and

the solvent. The potential energy mimics as accurately as possible the physical forces between atoms. To obtain the dynamics Newton's equations of motion are numerically integrated. This allows in principle to get the complete folding pathway. Clearly, the force fields used for MD simulations are very approximative since all quantum mechanical effects are captured in heuristic classical energy terms. This questions the validity of such simulations in general. However, it is well possible to get meaningful and justified results. If experimental data - they are always averages over ensembles - can be reproduced by simulation, it is reasonable to assume that the errors of the force field for the particular system under investigation are not dominant. In such a case the detailed analysis of the atomic motions can provide valuable insight in the mechanism of how the protein achieves structural changes.

The ergodicity theorem allows to introduce a simplification of great practical value into MD simulations. It implies that rather than simulating an ensemble of proteins it is sufficient to simulate one single protein for an appropriate time length. Nevertheless, MD simulations are confronted with the problem of sampling efficiency in order to provide statistically significant results. The heuristic energy functions used in MD simulations are much more detailed than the ones used for lattice models which slows down calculations significantly. As a consequence, several techniques have been developed to overcome the sampling barrier.

A major reason for unsatisfactory sampling in MD simulations is solvation. Solvation is crucial to protein structure and function. Hence, it can not be neglected in MD simulations. Although a 1 $\mu$s simulation in explicit water of the villin headpiece has been reported [28], the fact remains that 1 $\mu$s is still a rather short time for many biological processes, and that it is still an exceptional length for explicit water MD simulations. Clearly, the detailed motions of solvation molecules are of interest only in very sporadic occasions. It is therefore particularly frustrating that about 90% of the computation time is spent on the exact calculation of the solvent molecules. The most powerful approach to avoid this problem is to eliminate the solvent degrees of freedom by incorporating all solvent effects into a mean solvation term. This typically reduces the degrees of freedom by a factor of 10 to 15 and has the potential to speed up the simulations by a similar amount. The mean solvation term is added to the vacuo potential energy function. Ideally, it provides a system that is thermodynamically equivalent to the explicit treatment. However, efficient evaluation of a mean solvation term is not a trivial task. Several kinds of so-called implicit solvation models have been developed. Current work includes extensions to take a membrane environment into account. Common to most implicit solvation models is the continuum approximation. It treats the solvent as a dielectric medium

with a high dielectric constant. The protein region is assigned a low dielectric constant and otherwise left unchanged. The charge distribution is given by the partial charges of the protein. For more details, see section 3.3.2.

Another way to avoid the sampling barrier is to study unfolding rather than folding. In this case the native state is the starting structure and unfolding is achieved by either high temperature [29, 30, 31, 32, 33, 34] or a biasing potential. The validity of such simulations has been questioned [35, 36] since they are either not performed under physiological conditions or use an unphysical potential energy. Nevertheless, many interesting insights were gained for the protein folding process. Li and Daggett, for instance, were able to show that the fraction of native contacts in the transition state for a series of residues correlates reasonably well with the corresponding experimental $\phi$-values [33]. Brooks and coworkers demonstrated that the free energy surfaces for a few small proteins depend strongly on the native state topology [37, 38, 39], in agreement with the conclusion mentioned in section 2.2.

A further approach to ease the sampling barrier is to restrict explicit MD simulations to small peptides. More modern approaches take advantage of parallel computing. Replica exchange methods [40] and distributed computing [41] are very active areas of research. However, distributed computing methods need further development [42].

## 2.4 Current View of Protein Folding

From the theoretical point of view proteins are defined as polymers of amino acids that fulfill two requirements: the thermodynamic requirement and the kinetic requirement. The thermodynamic requirement postulates that under physiological conditions the stable native state of a protein corresponds to its global free energy minimum. The kinetic requirement postulates that the protein folds to the native state within a time short enough to be biologically meaningful. Consequently, there are two aspects to the protein folding problem. Given the sequence of a protein, the first aspect is to predict its native state, i.e., to predict its global free energy minimum (see [43] for a review). The second one is to elucidate the mechanism by which the protein finds its native state (see [44] for a review).

The thermodynamic requirement implies that the global free energy minimum be pronounced in units of the roughness of the free energy landscape under physiological conditions. This assures a thermodynamically stable native state. Otherwise the protein would, for instance, fluctuate between the native state and misfolded or partially unfolded conformations. Indeed, this is a common consequence of muta-

tions that are often followed by disease. A more precise statement of the implication of the thermodynamic requirement can be formulated. Let the roughness of the free energy landscape be characterized by $\Delta E$ and the typical energy gap between the native state and other partially misfolded configurations by $\delta E$. Then $\frac{\delta E}{\Delta E}$ must be large as a consequence of the thermodynamic requirement. Since any change in free energy is given by $\Delta F = \Delta H - T\Delta S$ and the folded state is a highly ordered state, it is reasonable to assume that the global free energy minium corresponds to the global potential energy minimum. This has been embodied in the principle of minimal frustration: Native contacts must on average be more favorable than non-native contacts.

The thermodynamic requirement says nothing about the folding time. Folding time is related to rate limiting steps, i.e., energy barriers. The kinetic requirement has implications for the highest free energy barrier the polypeptide chain has to cross in order to go from the unfolded to the folded state. Under physiological conditions this barrier must be small enough so that the protein can overcome it and become functional within a biologically meaningful time.

The current picture of folding of small proteins (less than about 100 residues) describes folding qualitatively as a downhill process on a free energy surface that in part has a funnel-like shape. Energy barriers are related to saddle points. This picture of protein folding includes but not restricts the possibility of certain preferred pathways and well defined sequences of events [45]. They were originally proposed by Levinthal but later found to be a too restricted view of protein folding.

## 2.5 Protein Misfolding

Protein misfolding is the most fundamental process in biology related to disease. There are two main causes for erroneous folding. Mutations are likely to change the thermodynamic and kinetic properties of a protein. As a consequence the protein may fluctuate between the native state and misfolded or partially unfolded conformations, or the native state is changed altogether. A second reason is that a freshly synthesized polypeptide chain is located in a cellular environment that is densely populated with macromolecules that have the potential to prevent the polypeptide chain to fold correctly. If as a consequence of a perturbed folding process the protein is only partially folded or fails to remain folded one possible implication is the formation of amyloid fibrils [46, 47]. Amyloid fibrils are insoluble aggregates that result from the self-assembly of partially unfolded proteins. They are found in at least twenty diseases for which no cure is available. Diseases caused by amyloid

deposits are known as amyloidosis and include Alzheimer's disease, Type II diabetes, Parkinson's disease, BSE, CJD, and fatal familial insomnia. Each of these diseases is related to the misfolding of a different protein. They do not share any sequence homology or common fold. Regardless of the native structure of the precursor proteins amyloid fibrils are polymers composed of proteins in cross $\beta$-sheet conformation [48, 49]. Proteins that are not known to form fibrils in vivo can do so under conditions where unfolded intermediates are well populated. This indicates that fibril formation can arise for most, if not all, polypeptide chains under certain conditions, and that nature has found ways to avoid fibrillogenic protein conformations. This, in turn, suggests that all ordered aggregation processes have common key elements. Therefore, the study of small and simplified systems that are able to form polymers in sheet configuration may provide valuable insight at atomic level into the pathologies related to protein deposits.

# Chapter 3

# Molecular Dynamics Simulations of Biological Molecules

## 3.1 Background

The fundamental physical laws that govern the interactions between atoms are well known. Given proper initial conditions the time evolution of a classical or quantum mechanical physical system is unique and can in principle be predicted. However, there are serious theoretical and practical obstacles to achieve this goal. From the theoretical point of view both the classical and quantum mechanical $n$-body problems are not integrable for $n \geq 3$. Only the 2-body problems (the earth-moon system in classical physics and the hydrogen atom in quantum mechanics, for instance) can be solved analytically. Numerical solutions provide an alternative approach. However, from the practical point of view the vast number of degrees of freedom pose a major challenge to numerical integration.

Clearly, the interactions between atoms should be treated on a quantum mechanical level. Apart from very small compounds this approach is too time consuming. Hence, most molecular dynamics (MD) simulations are based on classical physics. All atoms are idealized as van der Waals spheres. The potential energy function mimics the quantum mechanical effects by introducing heuristic energy terms. These terms include parameters that need to be calibrated. The standard approach is to first fit them to data obtained from quantum mechanical calculations on small compounds. Once the initial values are established, an iterative process follows to adjust the parameters so as to reproduce experimental results that were obtained for larger molecules. Classical potential energy functions can not simulate processes that involve significant changes in the quantum mechanical wave functions. Chemical reactions, i.e., breaking and building of bonds, can not be simulated with

such force fields. The dynamics in MD simulations is obtained by integration of Newton's classical equation of motion $F = m\ddot{x}$. This requires the potential energy function to have continuous first derivatives since $F = -\frac{\partial U}{\partial x}$. The explicit time integration used in nearly all MD software packages restricts the time step to the fastest motions that appear in the system. For proteins these are the oscillations of the hydrogen atoms.

## 3.2 Vacuo Force Field

The vacuo potential energy of CHARMM [50] is given by:

$$
\begin{aligned}
U^{vacuo} &= U_{bonded} + U_{non-bonded} \\
U_{bonded} &= U_b + U_\theta + U_\varphi + U_\omega \\
U_{non-bonded} &= U_{vdW} + U_{elec} \\
U_b &= \sum_b \frac{1}{2} k_b \left(r_b - r_{b_0}\right)^2 \\
U_\theta &= \sum_\theta \frac{1}{2} k_\theta \left(\theta - \theta_0\right)^2 \\
U_\varphi &= \sum_\varphi \frac{1}{2} k_\varphi \left(1 - \cos\left(n\varphi - \varphi_0\right)\right) \\
U_\omega &= \sum_\omega \frac{1}{2} k_\omega \left(\omega - \omega_0\right)^2 \\
U_{vdW} &= 4 \sum_{i<j} \epsilon_{ij} \left( \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right) \\
U_{elec} &= \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}
\end{aligned}
$$

The vacuo potential energy $U^{vacuo}$ is the sum of a bonded and a non-bonded part. The bonded part describes the covalent bonds. It mimics the quantum mechanics by introducing four heuristic potentials: the bond potential $U_b$, the bond angle potential $U_\theta$, the dihedral angle potential $U_\varphi$, and the improper angle potential $U_\omega$. The corresponding force constants are $k_b$, $k_\theta$, $k_\varphi$, and $k_\omega$. Since for the dihedral angle $\varphi$ several values correspond to the same optimal geometry a periodic function is chosen. For the other three terms only one value is best and a harmonic function is the simplest choice. The non-bonded part describes the non-covalent bonds. It mimics the quantum mechanical effects that arise between non-bonded atoms by introducing

the heuristic Lennard-Jones potential. The Lennard-Jones potential describes the van der Waals interactions due to fluctuating electron densities around the nuclei of atoms and the repulsion due to the Pauli exclusion principle. The electrostatic interaction energy between monopoles in vacuo is given by the Coulomb potential.

Furthermore, the CHARMM force field has several peculiarities. The most important ones are that the 1-2 and 1-3 pairs as well as certain 1-4 pairs, depending on the residue, are not included in the non-bonded energy terms. Additionally, the Coulomb interaction between 1-4 pairs is scaled and a cutoff is applied to all non-bonded energy terms.

## 3.3 Solvation

Solvation plays a crucial role in protein structure and function. Any form of life as we know it always occurs in an aqueous environment. Therefore, solvation can not be neglected in MD simulations in order to obtain a realistic picture of the processes going on in nature.

### 3.3.1 Explicit Solvation

In explicit solvation all solvent atoms are treated explicitly. Such an explicit treatment of solvation is accurate and simple in the framework of classical physics. The bottleneck is that the number of degrees of freedom and interaction centers is about an order of magnitude larger than for a vacuo system.

### 3.3.2 Implicit Solvation

The key concept of an implicit solvation model is to get rid of the solvent degrees of freedom. In this section the mathematics and physics underlying any implicit solvation model are presented. Although short and straightforward, it is helpful to see clearly what is common to all implicit solvation models, up to which point and in what sense they are rigorous, and where simplifications and approximations enter. The following remarks follow closely the derivation given in [51].

To clarify notational conventions two basic equations of thermodynamics are revisited. The partition function in the canonical ensemble of a physical system with Hamiltonian $H$ is defined by

$$Z = \int_{\mathbb{R}^n} dx^n e^{-\beta H(x)}$$

where $\beta = \frac{1}{k_B T}$ and $k_B$ denotes the Boltzmann constant and $T$ the temperature. All thermodynamic quantities can be derived from the partition function. The link between the microscopic states and the macroscopic quantities is given by the ensemble average. The ensemble average of a microscopic quantity $A$ that depends on the microscopic state $x$ is defined by

$$\langle A \rangle_H = \int_{\mathbb{R}^n} dx^n A(x) e^{-\beta H(x)}$$

Note that $\langle A \rangle_H$ is a macroscopic quantity that can be measured in an experiment.

Consider now a protein immersed in a solvent. Let $H$ denote the Hamiltonian of this physical system described by classical mechanics. An explicit representation is assumed which means that both solute and solvent atoms are included explicitly. Let $x$ denote the solute and $X$ the solvent degrees of freedom. In this case the Hamiltonian is additive in most empirical force force fields without polarization, which is true in particular for the CHARMM force field:

$$H(x, X) = H_{solute}(x) + H_{mixed}(x, X) + H_{solvent}(X)$$

$H_{solute}$ represents the intra-solute, $H_{mixed}$ the solute-solvent, and $H_{solvent}$ the intra-solvent interactions. The intra-solute Hamiltonian describes the vacuo force field. The partition function is given by

$$Z = \int_{\mathbb{R}^n \times \mathbb{R}^N} dx^n dX^N e^{-\beta H(x,X)}$$

One can formally integrate over the solvent degrees of freedom in order to get rid of them:

$$\begin{aligned} Z &= \int_{\mathbb{R}^n \times \mathbb{R}^N} dx^n dX^N e^{-\beta H(x,X)} && (3.1) \\ &= \int_{\mathbb{R}^n} dx^n e^{-\beta W(x)} && (3.2) \end{aligned}$$

where the definition

$$W(x) = -\frac{1}{\beta} \ln \left( \int_{\mathbb{R}^N} dX^N e^{-\beta H(x,X)} \right)$$

is used. The key implication of equations (3.1) and (3.2) is that the explicit system, described by the Hamiltonian $H$, and the implicit system, described by the

Hamiltonian $W$, are thermodynamically equivalent. Microscopic quantities differ since the phase spaces (the spaces of all possible configurations) of the explicit and implicit systems differ. All thermodynamic quantities - they are always averages over microscopic states - are identical. Note that $W$ depends only on the solute degrees of freedom. It is possible to write the Hamiltonian of the implicit system as a sum of the intra-solute Hamiltonian and the solvation free energy:

$$\begin{aligned} e^{-\beta W(x)} &= \int_{\mathbb{R}^N} dX^N e^{-\beta H(x,X)} \\ &= e^{-\beta H_{solute}(x)} \int_{\mathbb{R}^N} dX^N e^{-\beta H_{mixed}(x,X)} e^{-\beta H_{solvent}(X)} \\ &= e^{-\beta H_{solute}(x)} \left\langle e^{-\beta H_{mixed}} \right\rangle_{H_{solvent}} (x) \\ &= e^{-\beta H_{solute}(x)} e^{-\beta \triangle G^{slv}(x)} \\ &= e^{-\beta \left( H_{solute}(x) + \triangle G^{slv}(x) \right)} \end{aligned}$$

Here it is used that the ensemble average over the pure solvent of the interactions between the solute and the solvent is related to the excess chemical potential or the solvation free energy by

$$-\frac{1}{\beta} \ln \left\langle e^{-\beta H_{mixed}} \right\rangle_{H_{solvent}} (x) = \Delta G^{slv}(x)$$

Therfore the Hamiltonian of the implicit system is given by

$$W = H_{solute} + \Delta G^{slv}$$

This is a mathematically and physically rigorous result. The solvation free energy contains all solvent induced effects, namely:

- Solute-solvent electrostatic interactions
- Solute-solvent dispersion interactions (interactions between induced dipole moments, commonly referred to as van der Waals interactions)
- Entropy of the solvation due to the rearrangement of solvent atoms around the solute
- Clashes between solvent and solute atoms and between solvent atoms themselves, often referred to as friction

The quality of an implicit solvation model depends on how accurately $\Delta G^{slv}$ is approximated.

## 3.4 Continuum Electrostatics

A continuum electrostatic model describes the solute as a region with a low dielectric constant and a charge distribution defined by the partial charges of the solute atoms. The solvent is modeled as a region with a high dielectric constant. The electrostatic potential $\phi$ of a charge distribution $\rho$, given the dielectric function $\varepsilon$, is uniquely defined by the Poisson equation

$$\nabla\varepsilon\nabla\phi = -4\pi\rho$$

and the boundary conditions

$$\begin{aligned} \lim_{x\to\infty} x\phi\left(\overrightarrow{x}\right) &= \alpha \\ \lim_{x\to\infty} x^2\left(\nabla\phi\right)\left(\overrightarrow{x}\right) &= \beta \end{aligned}$$

The notation $x = |\overrightarrow{x}|$ is assumed and $\alpha$ and $\beta$ are finite real numbers. The appropriate electrostatic potential energy of the system is $U = \frac{1}{2}\int_{\mathbb{R}^3}\rho\phi$. For such a two-dielectric model as described above let $\varepsilon_{in}$ denote the dielectric constant of the macromolecular area and $\varepsilon_{out}$ the dielectric constant of the solvent area . The Poisson equation simplifies to

$$\varepsilon_{in}\Delta\phi = -4\pi\rho \tag{3.3}$$

$$\Delta\phi = 0 \tag{3.4}$$

where the first equation applies to the region of the macromolecule and the second equation to the region of the solvent. Due to the existence of a discontinuity surface additional boundary conditions are necessary to solve the Poisson equation. For points near the surface one has

$$\phi_{in} = \phi_{out} \tag{3.5}$$

$$\varepsilon_{in}\frac{\partial\phi_{in}}{\partial\overrightarrow{n}} = \varepsilon_{out}\frac{\partial\phi_{out}}{\partial\overrightarrow{n}} \tag{3.6}$$

where $\frac{\partial}{\partial\overrightarrow{n}}$ denotes the derivative in a direction perpendicular to the discontinuity surface. Note that if $\varepsilon_{in} = \varepsilon_{out}$ there is no discontinuity and the boundary conditions (3.5) and (3.6) follow from equations (3.3) and (3.4).

To obtain the electrostatic contribution to the solvation free energy one needs both the electrostatic potentials of the macromolecule in vacuo and immersed in solution. Let $\phi^m$ denote the electrostatic potential in vacuo, obtained by setting

$\varepsilon_{in} = \varepsilon_{out} = \varepsilon_m$. Analogously, let $\phi^s$ be the electrostatic potential with solvent present, obtained by setting $\varepsilon_{in} = \varepsilon_m$ and $\varepsilon_{out} = \varepsilon_s$. Typical values for the dielectric constant $\varepsilon_m$ of the macromolecule are 1, 2, or 4. Commonly used values for the dielectric constant $\varepsilon_s$ of the solvent are 78.5 and 80. The solvation free energy can be split into a non-polar and polar part:

$$\begin{aligned}
\Delta G^{slv} &= \Delta G^{non-polar} + \Delta G^{polar} \\
&= \Delta G^{non-polar} + \frac{1}{2}\int_{\mathbb{R}^3} \rho\,(\phi^s - \phi^m) \\
&= \Delta G^{non-polar} + \sum_{1\le i<j\le n} \int_{\mathbb{R}^3} \rho_i\left(\phi_j^s - \phi_j^m\right) + \sum_{i=1}^{n} \frac{1}{2}\int_{\mathbb{R}^3} \rho_i\,(\phi_i^s - \phi_i^m)
\end{aligned}$$

where $n$ denotes the total number of atoms in the system. There is a special case that can be solved analytically. Assume that each atom is treated as a hard sphere with a spherically symmetric charge distribution that is located on its surface. Assume furthermore that the atoms are separated by large distances, i.e., $r_i^{vdW} \ll r_{ij}$ where $r_{ij} = |\overrightarrow{x}_i - \overrightarrow{x}_j|$ and $r_i^{vdW}$ is the van der Waals radius. In this case one gets:

$$\Delta G^{slv} = \Delta G^{non-polar} - \sum_{1\le i<j\le n} \tau \frac{q_i q_j}{r_{ij}} - \sum_{i=1}^{n} \tau \frac{q_i^2}{2r_i^{vdW}} \tag{3.7}$$

where the abbreviation $\tau = \frac{1}{\varepsilon_m} - \frac{1}{\varepsilon_s}$ has been introduced.

## 3.5 The SASA Model

Equation (3.7) is the starting point for the SASA model. The aim is to make appropriate substitutions so as to obtain a formula valid also for close atoms. Three ingredients characterize the SASA model. The first one is the introduction of a distance dependent dielectric function to account for the screening effect of the solvation. The second one is the assumption that for a single atom, the solvation energy and the contribution to the non-polar part of the solvation free energy of the macromolecule are proportional to the solvent accessible surface area of the atom. These ideas suggest the following two substitutions in equation (3.7):

$$\begin{aligned}
-\sum_{1\le i<j\le n} \tau \frac{q_i q_j}{r_{ij}} &\longmapsto -\sum_{1\le i<j\le n} \left(\frac{1}{\varepsilon_m} - \frac{1}{2r_{ij}}\right) \frac{q_i q_j}{r_{ij}} \\
\Delta G^{non-polar} - \sum_{i=1}^{n} \tau \frac{q_i^2}{2r_i^{vdW}} &\longmapsto \sum_{i=1}^{n} \sigma_{T(i)} A_i
\end{aligned}$$

The $\{\sigma\}$ are the proportionality constants, and $A_i$ denotes the solvent accessible surface area of atom $i$. Altogether the solvation energy in the SASA model is given by

$$\Delta G^{slv,SASA} = -\sum_{1\leq i<j\leq n}\left(\frac{1}{\varepsilon_m}-\frac{1}{2r_{ij}}\right)\frac{q_iq_j}{r_{ij}}+\sum_{i=1}^{n}\sigma_{T(i)}A_i$$

The third ingredient is to use neutralized charges for the ionic side-chains [51] to further increase the screening effect. The number of solvation parameters $\sigma$ is restricted to two in the SASA model: one for carbon and sulfur atoms ($\sigma_{C,S} > 0$) and one for nitrogen and oxygen atoms ($\sigma_{N,O} < 0$). For hydrogen atoms $\sigma$ is set to 0. Deriving values by an optimization procedure [52] yields $\sigma_{C,S} = 0.012$ kcal/molÅ$^2$ and $\sigma_{N,O} = -0.060$ kcal/molÅ$^2$. The solvent accessible surface area is approximated by a fast method described in section 3.6. The SASA model is fully analytical and has been implemented in the official CHARMM version as part of this thesis. It has been applied in several studies [53, 54, 55, 56, 57].

## 3.6 Solvent Accessible Surface Area

The solvent accessible surface area in the SASA model is approximated by a probabilistic approach. It is based on an idea originally proposed by Wodak and Janin [58]. Let $A_i$ be the accessible surface of atom $i$ and let $r_i$ and $S_i$ denote the radius and area of its first solvation shell, respectively. The accessible surface removed from atom $i$ due to the overlap with atom $j$ is given by

$$b_{ij} = \begin{cases} \pi r_i\left(r_j + r_i - d\right)\left(1+\frac{r_j-r_i}{d}\right) & r_i + r_j < d \\ 0 & r_i + r_j \geq d \end{cases} \tag{3.8}$$

for an atomic separation of $d$. Note that equation (3.8) only holds if the center of atom $i$ is not within atom $j$ and vice versa, i.e., $d \geq r_i$ and $d \geq r_j$. Suppose now that a spot on the solvation shell of atom $i$ is chosen randomly. The probability to hit a part on the shell that overlaps with atom $j$ is $\frac{b_{ij}}{S_i}$. Consequently, the probability to hit a part on the shell that does not overlap with atom $j$ is $1-\frac{b_{ij}}{S_i}$. The product of these individual probabilities for all pairs that contain atom $i$ is the probability to hit a part on the shell of atom $i$ that does not overlap with any other atom. In other words, this product is the probability to hit the accessible surface of atom $i$ which is equal to $\frac{A_i}{S_i}$:

$$\frac{A_i}{S_i} = \prod_{j=1,\ldots,n,j\neq i}\left(1-\frac{b_{ij}}{S_i}\right) \tag{3.9}$$

However, this is only true under the assumption that all the atoms are distributed randomly which is not the case. Instead of elaborating into sophisticated probability calculations, Hasel and coworkers followed a pragmatic approach and parameterized equation (3.9) by including two sets of parameters, $\{u_t\}$ and $\{p_c\}$ [59]:

$$A_i = S_i \prod_{j=1,\ldots,n,j\neq i} \left(1 - \frac{u_{T(i)} p_{C(i,j)} b_{ij}}{S_i}\right) \tag{3.10}$$

$\{u_t\}$ are parameters depending on the atom type $t = T(i)$ of atom $i$ and correct for systematic errors primarily due to hybridization. $\{p_c\}$ are parameters that depend on the connectivity $c = C(i,j)$ of the atom pair $(i,j)$ and distinguish bound atoms from more distant ones. The parameters $\{u_t\}$ and $\{p_c\}$ are subject to the condition $u_t p_c < 1$ for all $t$ and $c$. A total of approximately 270 small molecules are used in [59] to determine the optimal parameters. Consequently, this parameter set is only suited for small peptides. For proteins a lot of inner cavities are produced and the discrimination between buried atoms and atoms on the surface is poor. Therefore, a reparameterization of equation (3.10) was done. This is described in the following.

To derive the new parameters a set of diverse structures is required. For this purpose ten proteins (1a2p [first chain], 1bpi, 1crn, 1hdn, 1pgb, 1pht, 1ubq, 2ci2, 2ptl, and beta3s) are chosen and subjected to high temperature unfolding simulations at 450 K for 20 ns with an implicit solvation model [52]. From each trajectory two conformations are chosen: the native state and a significantly unfolded one. The final set consists of twenty conformations that make up a total of 12'684 atoms. For each atom the exact solvent accessible surface area $A_i^{exact}$ is calculated with the analytical method of Lee and Richards in CHARMM. To obtain new parameters for equation (3.10) the fitness function

$$f = \sqrt{\frac{\sum_{i=1}^{n} (A_i^{apr} - A_i^{exact})^2}{n}}$$

is minimized by a particle swarm optimization algorithm [60]. The number of atoms is denoted by $n$ and the approximated surface of atom $i$ by $A_i^{apr}$. The new and old parameter sets are given in table 3.1 and 3.2. A comparison of their performance is shown in figures 3.1. The key improvement is demonstrated in figure 3.2. Clearly, the microcavities within the protein are no longer present with the new parameter set but the surface atoms are still recognized as such. To achieve this result it is crucial to include atoms of folded and unfolded structures in the data set that is used for the optimization.

| | SASA 1.0 | | SASA 2.0 | |
|---|---|---|---|---|
| Atom type | $r$ | $u$ | $r$ | $u$ |
| H | 1.100 | 1.128 | 0.100 | 0.989 |
| HC | 1.100 | 1.128 | 0.491 | 1.778 |
| HA | - | - | - | - |
| HT | - | - | - | - |
| LP | - | - | - | - |
| CT | - | - | - | - |
| C | 1.720 | 1.554 | 1.369 | 1.304 |
| CH1E | 1.800 | 1.276 | 1.896 | 1.316 |
| CH2E | 1.900 | 1.045 | 2.286 | 1.183 |
| CH3E | 2.000 | 0.880 | 2.651 | 1.083 |
| CR1E | 1.800 | 1.073 | 2.076 | 1.134 |
| CM | - | - | - | - |
| N | 1.550 | 1.028 | 0.324 | 1.130 |
| NR | 1.550 | 1.028 | 1.391 | 1.511 |
| NP | - | - | - | - |
| NH1 | 1.550 | 1.028 | 0.100 | 1.593 |
| NH2 | 1.600 | 1.215 | 1.470 | 1.262 |
| NH3 | 1.600 | 1.215 | 1.554 | 1.187 |
| NC2 | 1.600 | 1.215 | 1.552 | 1.054 |
| O | 1.500 | 0.926 | 1.682 | 1.036 |
| OC | 1.700 | 0.922 | 1.652 | 1.140 |
| OH1 | 1.520 | 1.080 | 1.549 | 1.097 |
| OH2 | - | - | - | - |
| OM | - | - | - | - |
| OT | - | - | - | - |
| OS | - | - | - | - |
| S | 1.800 | 1.121 | 2.189 | 1.680 |
| SH1E | 1.800 | 1.121 | 1.878 | 0.907 |
| FE | - | - | - | - |
| CR | 1.720 | 1.554 | 1.579 | 1.276 |

Table 3.1: Original (SASA 1.0) and new (SASA 2.0) surface parameters. For each atom type in the CHARMM parameter set 19, the optimized van der Waals radius $r$ (only used for surface calculations) and atom type parameter $u$ is shown. The unit of $r$ is Å.

| SASA 1.0 | | SASA 2.0 | |
|---|---|---|---|
| $p_{12}$ | $p_{13}$ | $p_{12}$ | $p_{13}$ |
| 0.8875 | 0.3516 | 0.3430 | 0.5074 |

Table 3.2: Original (SASA 1.0) and new (SASA 2.0) connectivity parameters. $p_{12}$ is for bonded atoms (1-2 pairs), $p_{13}$ for all other pairs.

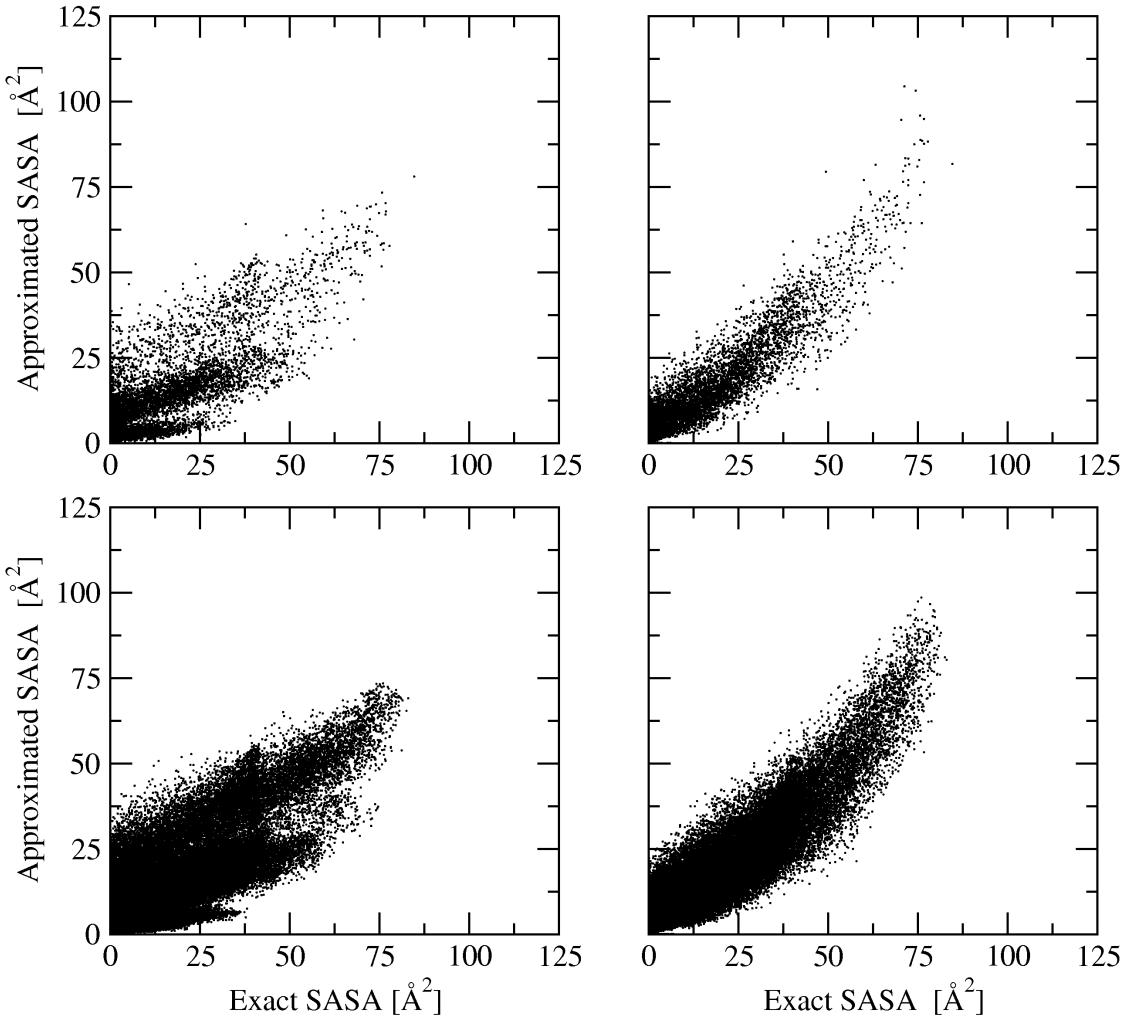


Figure 3.1: Top row: Solvent accessible surface areas of the 12684 atoms used in the optimization procedure. On the left, the areas calculated with the old parameters are shown. On the right, the new parameter set is applied. Bottom row: Solvent accessible surface areas of all atoms of 200 conformations of 1a2p along a high temperature unfolding trajectory. On the left, the areas calculated with the old parameters are shown. On the right, the new parameter set is applied.

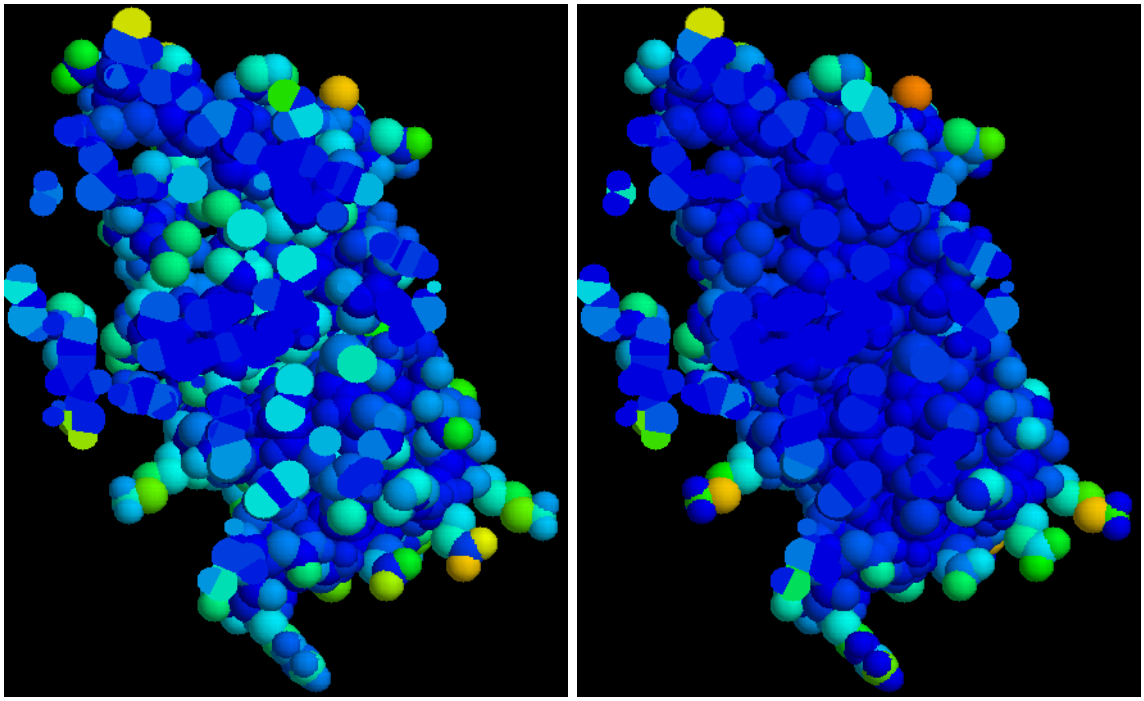

Figure 3.2: A slice through the middle of 1dcq is shown. The colors correspond to the atomic solvent accessible surface areas. Deep blue means an accessible area of approximately zero, and the lighter the color, the larger the accessible surface area. The picture on the left is created with the original parameters set, the picture on the right with the new one.

## 3.7 Generalized Born

The most reliable approximation to the solvation free energy of a collection of atoms surrounded by an arbitrary shape of the dielectric boundary is given by the Generalized Born formula that is obtained by performing the substitution

$$r_{ij} \mapsto \sqrt{r_{ij}^2 + R_i R_j e^{-\frac{1}{4}\frac{r_{ij}^2}{R_i R_j}}}$$

in equation (3.7) to yield

$$\Delta G^{slv,GB} = -\frac{1}{2}\sum_{i,j=1}^{N} \tau \frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i R_j e^{-\frac{1}{4}\frac{r_{ij}^2}{R_i R_j}}}} \tag{3.11}$$

where $R_i$ denotes the effective Born radius of atom $i$ and the convention $r_{ii} = 0$ is assumed. The effective Born radius is defined by

$$R_i = -\tau \frac{q_i^2}{2\Delta G_i^{slv}}$$

and $\Delta G_i^{slv}$ denotes the solvation energy of atom $i$. (This is the solvation energy of the macromolecule when all charges except the one of atom $i$ are deleted.) $\Delta G_i^{slv}$ is obtained, for instance, by numerical solution of the Poisson equation. Clearly, at this point the continuum approximation does not yet give a computational advantage over an explicit treatment. The key observation is that equation (3.11) yields very accurate results even if $\Delta G_i^{slv}$ is not perfect [61]. Therefore the problem reduces to the efficient and accurate determination of the effective Born radii or the atomic solvation energies. This is where the most recent developments of GB models have taken place. The standard approach is as follows. Firstly, the so called Coulomb field approximation (CFA) is assumed. In the CFA the dielectric displacement $\overrightarrow{D}_i$ for each atom $i$ is calculated by assuming that the dielectric boundary is spherical and that atom $i$ lies at the center of this sphere. As a consequence an analytical expression for the effective Born radii can be derived. This expression is evaluated by either a volume or a surface integral formulation. Recently, corrections to the CFA have been suggested and shown to greatly increase the accuracy of the effective Born radii [62].

The FACTS model (see chapter 5) differs significantly in several ways from the standard GB approach. The FACTS model originated from the AEI model [63] that allows to calculate screened interaction energies but not solvation energies. Firstly,

no integrations are performed in the FACTS model which results in a substantial speed advantage compared to integral based implementations. Secondly, no CFA is necessary which helps to increase accuracy. Thirdly, the determination of the dielectric boundary is only required for deriving the finite difference Poisson (fdP) reference data that are used to optimize the parameters of the FACTS model. After that, and in particular during MD simulations and for single point energy evaluations, no computation of the dielectric boundary is performed any more. This again facilitates a fast algorithm. In fact, the FACTS model is only about four times slower (including the calculation of derivatives) than MD simulations performed in vacuo. Single point energy evaluations that do not require derivatives are only marginally slower than plain in vacuo computations. Thus the FACTS model is significantly faster than the fastest previously published GB implementations while at the same time its accuracy is similar to the accuracy of most accurate GB implementations.

# Bibliography


[1] Karniadakis, G. E. and Kirby, R. M., *Parallel Scientific Computing in C++ and MPI: A Seamless Approach to Parallel Algorithms and their Implementation*, Cambridge University Press, 2003.

[2] Anfinsen, C. B., *Science*, 1973, **181**, 223–230.

[3] Creighton, T. E., *Proteins*, Freeman, 1984.

[4] Fersht, A., *Structure and Mechanism in Protein Science*, W. H. Freeman and Company, New York NY, 1999.

[5] Gruebele, M., *Annu. Rev. Phys. Chem.*, 1999, **50**, 485–516.

[6] Dill, K. A. and Chan, H. S., *Nature Struct. Biol.*, 1997, **4**, 10–19.

[7] Schindler, T.; Herrler, M.; Marahiel, M. A. and Schmid, F. X., *Nature Struct. Biol.*, 1995, **2**, 664–673.

[8] Jones, C. M.; Henry, E. R.; Hu, Y.; Chan, C.; Luck, S. D.; Bhuyan, A.; Roder, H.; Hofrichter, J. and Eaton, W. A., *Proc. Natl. Acad. Sci. USA*, 1993, **90**, 11860–11864.

[9] Nolting, B.; Golbik, R. and Fersht, A. R., *Proc. Natl. Acad. Sci. USA*, 1995, **92**, 10668–10672.

[10] Thompson, P. A.; Eaton, W. A. and Hofrichter, J., *Biochemistry*, 1997, **36**, 9200–9210.

[11] Briggs, M. S. and Roder, H., *Proc. Natl. Acad. Sci. USA*, 1989, **92**, 2017–2021.

[12] Huang, G. S. and Oas, T. G., *Proc. Natl. Acad. Sci. USA*, 1995, **92**, 6878–6882.

[13] Yang, J. T.; Wu, C. S. and Martinez, H. M., *Methods Enzymol.*, 1986, **130**, 208–269.

[14] Callender, R. H.; Dyer, R. B.; Gilmanshin, R. and Woodruff, W. H., *Ann. Rev. of Phys. Chem.*, 1998, **49**, 173–202.

[15] Gilmanshin, R.; Williams, S.; Callender, R. H. Woodruff, W. H. and Dyer, R. B., *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 3709–3713.

[16] Shastry, M. C.; Luck, S. D. and Roder, H., *Biophys. J.*, 1998, **74**, 2714–2721.

[17] Gruebele, M.; Sabelko, J.; Ballew, R. and Ervin, J., *Acc. Chem. Res.*, 1998, **31**, 699–707.

[18] Lillo, M. P.; Szpikowska, B. K.; Mas, M. T.; Sutin, J. D. and Beechem, J. M., *Biochemistry*, 1997, **36**, 11273 –11281.

[19] Fersht, A. R., *Curr. Opin. Struct. Biol.*, 1995, **5**, 79–84.

[20] Matouschek, A.; Kellis, Jr., J. T.; Serrano, L. and Fersht, A. R., *Nature*, 1989, **340**, 122–126.

[21] Itzhaki, L. S.; Otzen, D. E. and Fersht, A. R., *J. Mol. Biol.*, 1995, **254**, 260–288.

[22] Grantcharova, V. P.; Riddle, D. S.; Santiago, J. V. and Baker, D., *Nature Struct. Biol.*, 1998, **5**, 714–720.

[23] Martinez, J. C.; Pisabarro, M. T. and Serrano, L., *Nature Struct. Biol.*, 1998, **5**, 721–729.

[24] Baker, D., *Nature*, 2000, **405**, 39–42.

[25] Šali, A.; Shakhnovich, E. and Karplus, M., *Nature*, 1994, **369**, 248–251.

[26] Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D. and Wolynes, P. G., *Proteins: Structure, Function, and Bioinformatics*, 1995, **21**, 167–195.

[27] Wolynes, P.; Onuchic, J. and Thirumalai, D., *Science*, 1995, **267**, 1619–1620.

[28] Duan, Y. and Kollman, P. A., *Science*, 1998, **282**, 740–744.

[29] Daggett, V. and Levitt, M., *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 5142–5146.

[30] Mark, A. E. and van Gunsteren, W. F., *Biochemistry*, 1992, **31**, 7745–7748.

[31] Caflisch, A. and Karplus, M., *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 1746–1750.

[32] Li, A. and Daggett, V., *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 10430–10434.

[33] Li, A. and Daggett, V., *J. Mol. Biol.*, 1996, **257**, 412–429.

[34] Lazaridis, T. and Karplus, M., *Science*, 1997, **278**, 1928–1931.

[35] Finkelstein, A. V., *Protein Engineering*, 1997, **10**, 843–845.

[36] Walser, R.; Mark, A. E. and van Gunsteren, W. F., *Biophys. J.*, 2000, **78**, 2752–2760.

[37] Boczko, E. M. and Brooks III, C. L., *Science*, 1995, **269**, 393–396.

[38] Sheinerman, F. and Brooks III, C., *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 1562–1567.

[39] Sheinerman, F. B. and Brooks III, C. L., *J. Mol. Biol.*, 1998, **278**, 439–456.

[40] Rao, F. and Caflisch, A., *J. Chem. Phys.*, 2003, **119**, 4035–4042.

[41] Larson, S. M.; Snow, C. D.; Shirts, M. and Pande, V. S.

[42] Paci, E.; Cavalli, A.; Vendruscolo, M. and Caflisch, A., *Proc. Natl. Acad. Sci. USA*, 2003, **100**, 8217–8222.

[43] Dobson, C. M.; Šali, A. and Karplus, M., *Angew. Chem. Int. Ed.*, 1998, **37**, 869–893.

[44] Moult, J.; Fidelis, K.; Zemla, A. and Hubbard, T., *Proteins: Structure, Function, and Genetics*, 2003, **53**, 334–339.

[45] Levinthal, C., *J. Chim. Phys.*, 1968, **65**, 44–45.

[46] Thomas, P.; Qu, B. and Pedersen, P., *Trends in Biochemical Sciences*, 1995, **20**, 456–459.

[47] Dobson, C. M., *Trends Biochem. Sci.*, 1999, **24**, 329–332.

[48] Blake, C. and Serpell, L., *Structure*, 1996, **4**, 989–998.

[49] Malinchik, S. B.; Inouye, H.; Szumowski, K. E. and Kirschner, D. A., *Biophys. J.*, 1998, **74**, 537–545.

[50] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S. and Karplus, M., *J. Comput. Chem.*, 1983, **4**, 187–217.

[51] Lazaridis, T. and Karplus, M., *Proteins: Structure, Function, and Bioinformatics*, 1999, **35**, 133–152.

[52] Ferrara, P.; Apostolakis, J. and Caflisch, A., *Proteins: Structure, Function, and Bioinformatics*, 2002, **46**, 24–33.

[53] Ferrara, P. and Caflisch, A., *Proc. Natl. Acad. Sci. USA*, 2000, **97**, 10780–10785.

[54] Ferrara, P. and Caflisch, A., *J. Mol. Biol.*, 2001, **306**, 837–850.

[55] Hiltpold, A.; Ferrara, P.; Gsponer, J. and Caflisch, A., *J. Phys. Chem. B*, 2000, **104**, 10080–10086.

[56] Gsponer, J. and Caflisch, A., *J. Mol. Biol.*, 2001, **309**, 285–298.

[57] Gsponer, J. and Caflisch, A., *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 6719–6724.

[58] Wodak, S. J. and Janin, J., *Proc. Natl. Acad. Sci. USA*, 1980, **77**, 1736–1740.

[59] Hasel, W.; Hendrickson, T. F. and Still, W. C., *Tetrahedron Comput. Methodol.*, 1988, **1**, 103–116.

[60] Kennedy, J. and Eberhart, R. C., *Swarm Intelligence*, Morgan Kaufmann Publishers, 2001.

[61] Onufriev, A.; Bashford, D. and Case, D. A., *J. Comput. Chem.*, 2002, **23**, 1297–1304.

[62] Lee, M. S.; Salsbury, F. R. and Brooks III, C. L., *J. Chem. Phys.*, 2002, **116**, 10606–10614.

[63] Haberthür, U.; Majeux, N.; Werner, P. and Caflisch, A., *J. Comput. Chem.*, 2003, **24**, 1936–1949.

# Part II

# Publications

# Chapter 4

# Efficient Evaluation of the Effective Dielectric Function of a Macromolecule in Aqueous Solution

# Efficient Evaluation of the Effective Dielectric Function of a Macromolecule in Aqueous Solution

**URS HABERTHÜR, NICOLAS MAJEUX, PHILIPP WERNER, AMEDEO CAFLISCH**
*Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland*

*Received 17 February 2003; Accepted 17 April 2003*

**Abstract:** We propose an analytical approach to calculate the effective dielectric function of proteins in aqueous solution. The screening effect if quantified by a measure of enclosure which is based on the distribution of solute atomic volumes around a pair of charges in a macromolecule. For protein conformations that vary significantly in size and shape, a comparison with finite difference Poisson calculations shows that pair interaction energies, their sums and solvation energies are well reproduced. The approach rivals the speed of simple distance dependent dielectric functions and the accuracy of the generalized Born model.

 J Comput Chem 24: 1936–1949, 2003

**Key words:** implicit solvent; continuum dielectric; analytical electrostatic interaction (AEI); generalized Born approach; solvation energy; effective dielectric constant

## Introduction

Incorporating solvent effects in molecular dynamics (MD) and Monte Carlo simulations is of key importance to quantitatively understand chemical and physical properties of biomolecular processes. Accurate electrostatic energies of proteins in an aqueous environment are one indispensable component to discriminate between native and non-native conformations. An exact evaluation of electrostatic energies considers the interactions among all possible solute-solute, solute-solvent, and solvent-solvent pairs of charges. However, this is computationally expensive for macromolecules. Continuum dielectric approximations offer a more tractable approach.[1–5] The essential concept in continuum models is to represent the solvent by a high dielectric medium, which eliminates the solvent degrees of freedom, and to describe the macromolecule as a region with a low dielectric constant and a spatial charge distribution. The Poisson equation provides an exact description of such a system. The increase in computation speed for a finite difference solution of the Poisson equation,[6–9] with respect to an explicit treatment of the solvent, is remarkable, but still not enough for effective utilization in computer simulations of macromolecules. The generalized Born (GB) model was introduced to facilitate an efficient evaluation of continuum electrostatic energies.[10] It provides accurate energetics and the most efficient implementations are between five to ten times slower than *in vacuo* simulations.[11–13] The essential element of the GB approach is the calculation of an effective Born radius for each atom in the system, which is a measure of how deeply the atom is buried inside the protein. This information is combined in a heuristic way to obtain a correction to the Coulomb law for each atom pair.[10] For the integration of energy density, necessary to obtain the effective Born radii, both numerical[10,11,13] and analytical[13–16] implementations exist. The former are more accurate but slower than the latter.[13] Moreover, analytical derivatives that are required for MD simulations are not given by numerical implementations.

For efficiency reasons empirical dielectric screening functions are the most common choice in MD simulations with implicit solvent. One kind of solvation model is based on the use of a dielectric function that depends linearly on the distance $r$ between two charges [$\varepsilon(r) = \alpha r$][17,18] or has a sigmoidal shape.[19–22] Although very fast, these options suffer from their inability to discriminate between buried and solvent exposed regions of a macromolecule and are therefore rather inaccurate. Recently, a distance- and exposure-dependent dielectric function was proposed.[23]

The aim of this article is to give an analytical approximation of the effective dielectric screening function that rivals the speed of $\varepsilon(r) = \alpha r$ and the accuracy of the GB model. A measure of

***Correspondence to:*** A. Caflisch

Contract/grant sponsor: Swiss National Competence Center in Structural Biology (NCCR)

Contract/grant sponsor: Swiss National Science Foundation; contract/grant number: 31-64968.01 (A.C.)

enclosure that focuses directly on atom pairs and their neighborhoods is introduced. It provides an approximate description of where the atom pair is located with respect to the bulk of the macromolecule and the solvent. A fit to effective dielectric constants derived from finite difference Poisson (fdP) energies for a set of several protein structures supplies analytical functions with continuous derivatives. The question of transferability and predictive power of the model presented here, henceforth called the analytical electrostatic interaction (AEI) model, is addressed by dividing the set of protein structures into several training and test sets. Various comparisons with electrostatic energies calculated by fdP, the GB approach,[13] and the sigmoidal distance dependent dielectric (SGM) model[20] are given. Finally, the physical relevance of the measure of enclosure is analyzed by comparing AEI with fdP solvation energies.

## Methods

### *AEI Model*

#### *Theory*

Consider a macromolecule in a fixed configuration immersed in a polar solvent with zero ionic strength. The Poisson equation

$$\nabla[\varepsilon(\vec{x})\nabla\phi(\vec{x})] = -4\pi\rho(\vec{x}) \tag{1}$$

defines the electrostatic potential $\phi$ given the dielectric function $\varepsilon$ and the charge density $\rho$. In the continuum approximation used in all following calculations $\varepsilon(\vec{x}) = \varepsilon_m$ for the region of the macromolecule and $\varepsilon(\vec{x}) = \varepsilon_s$ for the region of the solvent. The effective dielectric constant $\varepsilon_{ij}^{\mathrm{fdP,eff}}$ for each pair of atoms $i$ and $j$ is defined such that if substituted into the Coulomb law the same electrostatic interaction energy results as when solving the Poisson equation:

$$q_i q_j \phi_i^{\mathrm{fdP}}(\vec{x}_j) =: \frac{q_i q_j}{r_{ij}\varepsilon_{ij}^{\mathrm{fdP,eff}}} \Leftrightarrow \varepsilon_{ij}^{\mathrm{fdP,eff}} := \frac{1}{r_{ij}\phi_i^{\mathrm{fdP}}(\vec{x}_j)} \tag{2}$$

where $\phi_i^{\mathrm{fdP}}$ is the electrostatic potential of a unit charge at the position of atom $i$; $q_i$ and $q_j$ denote the charges of atoms $i$ and $j$, respectively; $\vec{x}_j$ represents the position of atom $j$, and $r_{ij}$ the distance between atoms $i$ and $j$. Note that $q_i q_j \phi_i^{\mathrm{fdP}}(\vec{x}_j)$ is the electrostatic interaction energy of the $(i, j)$ pair in the presence of solvent.

For an accurate approximation of $\varepsilon_{ij}^{\mathrm{fdP,eff}}$ it is necessary to discriminate between buried and solvent exposed atoms in the macromolecule. In the GB approach[10] the effective dielectric constant $\varepsilon_{ij}^{\mathrm{GB,eff}}$ is defined by

$$\frac{1}{\varepsilon_{ij}^{\mathrm{GB,eff}}} = \frac{1}{\varepsilon_m} - \left(\frac{1}{\varepsilon_m} - \frac{1}{\varepsilon_s}\right)\left(1 + \frac{R_i R_j}{r_{ij}^2} e^{-(1/4)(r_{ij}^2/R_i R_j)}\right)^{-1/2} \tag{3}$$

$$= \frac{1}{\varepsilon_m} - \left(\frac{1}{\varepsilon_m} - \frac{1}{\varepsilon_s}\right)\left(1 + \left(\frac{u_{ij}^{\mathrm{GB}}}{r_{ij}}\right)^2 e^{-(1/4)(r_{ij}/u_{ij}^{\mathrm{GB}})^2}\right)^{-1/2} = f\left(\frac{u_{ij}^{\mathrm{GB}}}{r_{ij}}\right) \tag{4}$$

where $R_i$ and $R_j$ denote the effective Born radii of atoms $i$ and $j$, respectively, $u_{ij}^{\mathrm{GB}} = \sqrt{R_i R_j}$, and the function $f$ is defined by $f(x) = 1/\varepsilon_m - (1/\varepsilon_m - 1/\varepsilon_s)/\sqrt{1 + x^2 e^{-1/4x^2}}$. The effective Born radius of an atom in the system is a measure of its enclosure. Consequently, the quantity $u_{ij}^{\mathrm{GB}}$ could be interpreted as a measure of enclosure of the $(i, j)$ atom pair: the larger $u_{ij}^{\mathrm{GB}}$ is the more buried the $(i, j)$ pair is. Because the calculation of the effective Born radii is the time consuming part in the GB approach and because this article is mainly concerned with interaction energies, we seek an alternative way to quantify the degree of enclosure of an atom pair in the macromolecule. We introduce a new and computationally efficient measure of enclosure $u_{ij}^{\mathrm{AEI}}$ and approximate $\varepsilon_{ij}^{\mathrm{fdP,eff}}$ in the same spirit as in the GB model by a function of $u_{ij}^{\mathrm{AEI}}/r_{ij}$, that is, $1/\varepsilon_{ij}^{\mathrm{fdP,eff}} \cong 1/\varepsilon_{ij}^{\mathrm{AEI,eff}} = g(u_{ij}^{\mathrm{AEI}}/r_{ij})$.

The present approach to calculate a measure of enclosure $u_{ij}^{\mathrm{AEI}}$ focuses on a finite region $\Omega_{ij}$ of space. This region is chosen around atoms $i$ and $j$, and is large enough to neglect effects on $\varepsilon_{ij}^{\mathrm{fdP,eff}}$ due to conformational changes outside $\Omega_{ij}$. The exact shape of this region is not important for the following arguments. One could, for instance, imagine a cylinder with an axis along the line joining atoms $i$ and $j$ or a sphere or an ellipsoid with its center somewhere between the two atoms. If only atoms $i$ and $j$ of the macromolecule were present within $\Omega_{ij}$, solving the Poisson equation would result in $\varepsilon_{ij}^{\mathrm{fdP,eff}} \cong \varepsilon_s$ and $u_{ij}^{\mathrm{AEI}}$ is required to be small. As more and more atoms are gradually added, $\varepsilon_{ij}^{\mathrm{fdP,eff}}$ decreases and $u_{ij}^{\mathrm{AEI}}$ increases in a complex way depending on where the additional atoms are placed. When all the solvent has finally been flushed out from $\Omega_{ij}$, solving the Poisson equation would result in $\varepsilon_{ij}^{\mathrm{fdP,eff}} \cong \varepsilon_m$ and $u_{ij}^{\mathrm{AEI}}$ reaches its maximum value. Intuitively, one expects that atoms located near or between charges $i$ and $j$ increase $u_{ij}^{\mathrm{AEI}}$ more than atoms located far from the $(i, j)$ pair because the closer an atom is placed to atoms $i$ and $j$, the more it influences the electric field at their positions.[24,25] Furthermore, adding a large atom is expected to increase $u_{ij}^{\mathrm{AEI}}$ more than adding a small one because more solvent is displaced.

The above arguments suggest quantification of the degree of enclosure of the $(i, j)$ atom pair by a function that depends on the sum of the van der Waals volumes within $\Omega_{ij}$, which are weighted according to their positions with respect to atoms $i$ and $j$. In the GB approach the measure of enclosure $u_{ij}^{\mathrm{GB}}$ is the square root of the product of the effective Born radii of atoms $i$ and $j$. In the AEI model $u_{ij}^{\mathrm{AEI}}$ is the square of a sum of weighted van der Waals volumes located around the atom pair. While there are many methods of calculating a weighted sum, the necessity for low computational costs eliminates most of them. For instance, it is not feasible to construct a cylinder around each atom pair and calculate a weighted sum within such a cylinder. We will only use quantities already available in the course of an MD simulation.

Two spheres of radius $r_{\mathrm{sphere}}$ with centers at the positions of atoms $i$ and $j$ define $\Omega_{ij}$. Let $A$ denote the set of all atoms with their centers within the sphere around atom $i$. Note that atom $i$ belongs to $A$. Let $B$ denote the corresponding set of atoms for the sphere around atom $j$. Let $v_k$ be the van der Waals volume of any atom $k$ and $N$ the total number of atoms of the macromolecule. Then $u_{ij}^{\mathrm{AEI}}$ is defined by the square of a sum of weighted van der Waals volumes of the atoms in $A$ and $B$ (see Fig. 1):

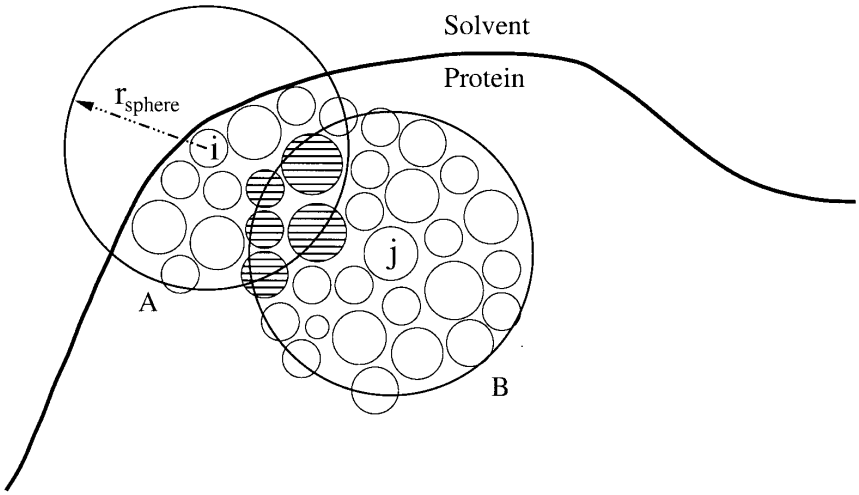


**Figure 1.** Schematic illustration for the calculation of the measure of enclosure $u_{ij}^{\mathrm{AEI}}$ in the case $r_{\mathrm{sphere}} < r_{ij} < 2r_{\mathrm{sphere}}$. The small circles describe protein atoms. The two large circles represent the spheres that define the neighborhood of atoms $i$ and $j$, which are taken into account to evaluate $u_{ij}^{\mathrm{AEI}}$. Atoms within the large spheres around atoms $i$ and $j$ constitute sets $A$ and $B$, respectively. The shaded circles represent the atoms in the intersection of $A$ and $B$. They are weighted with respect to the positions of both atoms $i$ and $j$. The atoms described by the small empty circles are only weighted with respect to either atom $i$ or atom $j$.

$$u_{ij}^{\mathrm{AEI}} := \left( \sum_{k=1}^{N} v_k \Theta_{ik} + \sum_{k=1}^{N} v_k \Theta_{jk} \right)^2 \tag{5}$$

$$= \left( \sum_{k \in A \setminus B} v_k \Theta_{ik} + \sum_{k \in B \setminus A} v_k \Theta_{jk} + \sum_{k \in A \cap B} v_k (\Theta_{ik} + \Theta_{jk}) \right)^2 \tag{6}$$

where the weighting function $\Theta_{ik}$ is defined by

$$\Theta_{ik} := \begin{cases} \left(1 - \left(\frac{r_{ik}}{r_{\mathrm{sphere}}}\right)^2\right)^2 & r_{ik} < r_{\mathrm{sphere}} \\ 0 & r_{ik} \geq r_{\mathrm{sphere}} \end{cases} \tag{7}$$

The first two sums in eq. (6) include all atoms in $A$ and $B$ that are not in the intersection of $A$ and $B$. The volumes of these atoms are weighted with respect to the position of either atom $i$ or atom $j$. The third sum in eq. (6) includes all atoms in the intersection of $A$ and $B$. The volumes of these atoms are weighted with respect to the positions of both atoms $i$ and $j$. The weighting function $\Theta_{ik}$ assures that the further away an atom $k$ is placed from atom $i$, the lower its weight and the less it contributes to the sum. Furthermore, the existence of continuous derivatives (required for MD simulations) is guaranteed. Note that $\Theta_{ik}$ is the shifting function that is commonly used in MD simulations to have zero Coulomb interaction energy at the cutoff.[26] Only if $r_{ij} < 2r_{\mathrm{sphere}}$ do the spheres around atoms $i$ and $j$ overlap. If $r_{ij} \geq 2r_{\mathrm{sphere}}$, each atom is only weighted with respect to the position of either atom $i$ or atom $j$, but in this case the distance $r_{ij}$ is large, thus the interaction energy is small and a more crude approximation justified. Following the arguments mentioned previously

$$\frac{1}{\varepsilon_{ij}^{\mathrm{fdP,eff}}} \cong g\left(\frac{u_{ij}^{\mathrm{AEI}}}{r_{ij}}\right) \tag{8}$$

and the function $g$ has to be chosen so as to approximate $1/\varepsilon_{ij}^{\mathrm{fdP,eff}}$ as accurately as possible. The calculation of the measure of enclosure $u_{ij}^{\mathrm{AEI}}$ [see eqs. (5) and (7)] can be performed very efficiently. Building a list of atoms within a sphere of radius $r_{\mathrm{sphere}}$ around each atom in the system is intrinsic to MD simulations. The same is true of the shifting function and its derivative. Because the derivation given in this section is heuristic, the approach is ultimately only justified if the results are satisfactory.

### *Determination of the Function g*

For a training set of structures the function $g$ in eq. (8) was determined by fitting it to the inverse of effective dielectric constants derived from fdP energies [see eq. (2)]. The performance was assessed by comparing interaction energies calculated by the AEI model and by fdP for the conformations in a test set. Of particular interest was whether or not good performance of both folded and unfolded states for peptides and larger proteins could be achieved.

An initial set of 29 proteins (23 single and six multichain proteins) of very different sizes and shapes was used. The struc-

**Table 1.** Definitions of the Seven Test Cases that Are Used to Perform Cross Correlations.

| | Training set | | Test set | |
|---|---|---|---|---|
| Test case | Type of structures | Number | Type of structures | Number |
| a | All | 52 | All | 52 |
| b | Randomly selected | 26 | The complementary set | 26 |
| c | Test set (b) | 26 | Training set (b) | 26 |
| d | Native | 29 | Unfolded | 23 |
| e | Unfolded | 23 | Native | 29 |
| f | Less than 70 amino acids | 27 | More than 70 amino acids | 25 |
| g | More than 70 amino acids | 25 | Less than 70 amino acids | 27 |

A test case consists of a training and a test set. For each test case the parameters of the AEI model are derived from the training set and used to calculate interaction energies for the structures in the test set. Apart from test case (a), the training and test sets are disjointed.

tures ranged in size from 11 (1cb3) to 347 (3pte) amino acids. The set included almost spherical geometries with no microcavities, as well as structures with internal cavities. 5hvp, for instance, is the HIV-1 aspartic proteinase in a complex with a peptidic ligand that was removed from the active site to obtain an internal cavity. To further diversify the set of structures with many different kinds of irregular shapes (cavities, open loops, etc.), the single chain proteins were subjected to high temperature unfolding simulations at 450 K for 20 ns with an implicit solvation model.[27] From each trajectory an unfolded conformation was selected and added to the initial set of structures. The average increase in the radius of gyration of the chosen conformations was 32% and their average $C_\alpha$-RMSD was 12.8 Å. The final set consisted of 52 conformations. All atoms (a total of 47,979 atoms) were assigned unit charges, and all pair interaction energies for every conformation in the set of the 52 conformations (a total of 39,041,961 pairs) were calculated by numerical (finite difference) solution of the Poisson equation.

The set of 52 conformations was divided in seven different ways into a training and a test set in order to perform cross correlations (Table 1). While cases (a), (b), and (c) addressed the convergence of the parameterization in general, cases (d) and (e) investigated how well the parameters extrapolate to different shapes, and cases (f) and (g) investigated how well the parameters extrapolate to different sizes. Note that apart from case (a), the training and test sets are disjointed.

Given a specific training set, $g(u_{ij}^{\mathrm{AEI}}/r_{ij})$ was fitted to $1/\varepsilon_{ij}^{\mathrm{fdP,eff}}$ rather than a function $\tilde{g}(u_{ij}^{\mathrm{AEI}}/r_{ij})$ fitted to $\varepsilon_{ij}^{\mathrm{fdP,eff}}$ in order to obtain accurate values for small effective dielectric constants (only these can result in high energies). Three different cases were distinguished: 1-2 pairs, 1-3 pairs, and all other pairs, that is

$$\frac{1}{\varepsilon_{ij}^{\mathrm{fdP,eff}}} \cong g_k\left(\frac{u_{ij}^{\mathrm{AEI}}}{r_{ij}}\right) \tag{9}$$

where $k = 1$ for 1-2 pairs, $k = 2$ for 1-3 pairs, and $k = 3$ for all remaining pairs. (A 1-2 pair consists of two covalently bonded atoms and a 1-3 pair of two atoms covalently bonded to a common atom.) For each of the three cases the range of the variable $u_{ij}^{\mathrm{AEI}}/r_{ij}$ was divided into 100 bins and the average of all $1/\varepsilon_{ij}^{\mathrm{fdP,eff}}$ values within a given bin was calculated. The functions $g_k$ were obtained by fitting analytical functions of the form of $f$ in eq. (4) to the average curves (see Appendix). Note that the functions $g_k$ have continuous derivatives.

The measure of enclosure $u_{ij}^{\mathrm{AEI}}$, defined by eqs. (5) and (7), and the analytical functions $g_k$ given in the Appendix, are the main results of this article and constitute the AEI model. They were used to calculate electrostatic interaction energies $E_{ij}$ for solute charges immersed in solvent by the formula

$$E_{ij} = 332\,\frac{q_i q_j}{r_{ij}}\,g_k\left(\frac{u_{ij}^{\mathrm{AEI}}}{r_{ij}}\right) \tag{10}$$

where the factor 332 was introduced to obtain values in kcal/mol. Note that only $g_3$ is relevant for MD simulations because the interaction energies of 1-2 and 1-3 pairs are accounted for in the bonding terms of the force fields. Results are also presented for 1-2 and 1-3 pairs to show that the approach is valid in general. The calculation of solvation energies within the framework of the AEI model is outlined in the section Solvation Energies.

### *Finite Difference Poisson*

The numerical (finite difference) solution of the Poisson equation was calculated with the PBEQ module[28] in CHARMM.[26] A grid spacing of 0.3 Å was used. (Some calculations with a grid spacing of 0.2 Å were also performed. Relative errors of interaction energies for a grid spacing of 0.3 Å with respect to a grid spacing of 0.2 Å are on average only about 0.55%.) The dielectric discontinuity surface was defined by the molecular surface. This is the surface spanned by the surface of a solvent probe sphere of radius 1.4 Å rolled over the van der Waals envelope of the atoms. The molecular volume was treated as a dielectric medium with a low dielectric constant $\varepsilon_m = 1$. Together with the spatial charge distribution of the macromolecule it represents the solute. The remaining space was treated as a dielectric medium with a high dielectric constant $\varepsilon_s = 78.5$ and represents the solvent. Solvation energies were calculated by subtracting the *vacuo* self-energy

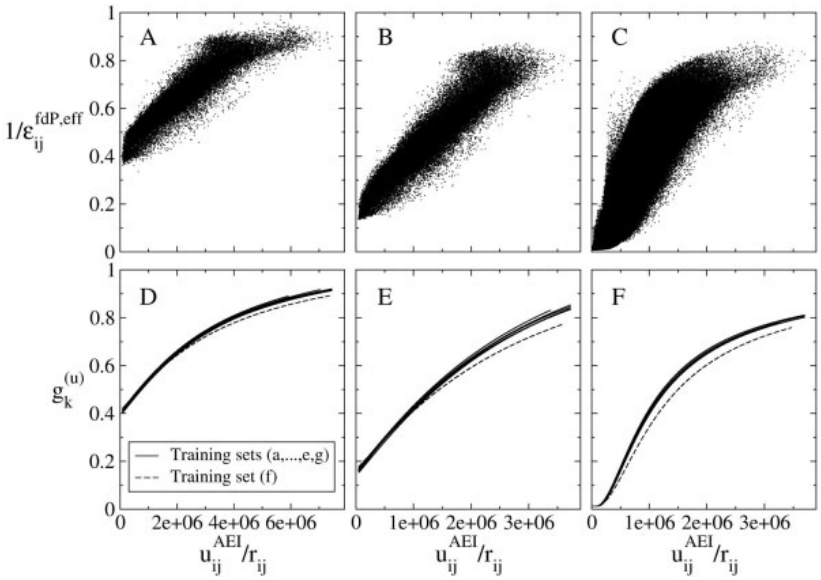


**Figure 2.** Top: inverse of fdP-derived effective dielectric constants against $u_{ij}^{AEI}/r_{ij}$ for 1-2 pairs (A), 1-3 pairs (B), and all remaining pairs (C) of all 52 conformations. Bottom: analytical functions resulting from the fits (see text) for 1-2 pairs (D), 1-3 pairs (E), and all remaining pairs (F). The solid lines represent the fits for training sets (a) to (e) and (g). The dotted lines represent the fits for training set (f).

($\varepsilon_m = 1$, $\varepsilon_s = 1$) from the self-energy in solution ($\varepsilon_m = 1$, $\varepsilon_s = 78.5$). Note that using $\varepsilon_m = 1$ (instead of $\varepsilon_m = 2$ or $\varepsilon_m = 4$, for instance) is the most stringent test for the accuracy of the AEI model. For single-point energy calculations (e.g., for ranking in ligand binding), $\varepsilon_m > 1$ would be more appropriate because it accounts for thermal fluctuations. As the AEI model is primarily aimed to be used in MD simulations, the more stringent test with $\varepsilon_m = 1$ was chosen for the present validation.

### *GB*

The GB calculations were performed with the analytical implementation of the GBMV module[13] in CHARMM. The dielectric discontinuity surface and the dielectric constants were defined as in the fdP calculations. Note that the analytical GBMV reproduces the Poisson-derived Born radii with an accuracy of about 2–4% and with a correlation of about 0.95.[13] In previous analytical implementations of the GB model there is cancellation of errors because the Coulomb approximation tends to overestimate the effective Born radii, whereas the analytical approximation of the energy density integration tends to underestimate them.[29]

### *SGM Function*

The SGM model is based on the sigmoidal function[19–21]:

$$\varepsilon^{SGM}(r_{ij}) = A + \frac{B}{1 + \alpha e^{-\beta r_{ij}}} \qquad (11)$$

which is similar to the one used recently.[22] The parameters $A$ and $B$ were determined by the conditions $\lim_{r_{ij}\to 0}\varepsilon^{SGM}(r_{ij}) = \varepsilon_m = 1$ and $\lim_{r_{ij}\to\infty}\varepsilon^{SGM}(r_{ij}) = \varepsilon_s = 78.5$. The two remaining parameters $\alpha$ and $\beta$ were determined by optimizing $\varepsilon^{SGM}$ for each of the 52 conformations (separately because of memory requirements) in such a way that fdP interaction energies were reproduced as accurately as possible. Finally, the values for $\alpha$ and $\beta$ were averaged over the 52 conformations. The resulting parameters were: $\alpha = 60.868$, $\beta = 0.317541$, and $A = -0.273247$, $B = 78.773247$. Note that $A$ and $B$ are such that if interaction energies are calculated by $E_{ij} = 332 q_1 q_2 / r_{ij} \varepsilon^{SGM}(r_{ij})$, the units are kcal/mol.

### *Parameter Set*

All calculations were performed using the van der Waals radii and partial charges of the CHARMM parameter set PARAM19.[26] For the fdP calculations and certain tests (see below) all atoms were assigned unit charges. The CHARMM parameter set PARAM19 treats hydrogens covalently bound to carbons implicitly and polar hydrogens explicitly.

## Results and Discussion

For all seven training sets (see Table 1) the functions $g_k$ with $k \in \{1, 2, 3\}$ are determined following the prescription in the section Determination of the Function $g$. They are denoted by $g_k^{(u)}$ with

**Table 2.** Cross Correlation Data for Pair Interaction Energies $E_{ij}$ Calculated by the AEI Model and by fdP.

| Test case | Correlation | Slope | RMSD |
|---|---|---|---|
| a | 0.976 | 1.033 (0.081) | 1.400 |
| b | 0.972 | 0.977 (0.082) | 1.383 |
| c | 0.980 | 1.085 (0.091) | 1.469 |
| d | 0.972 | 0.981 (0.054) | 1.072 |
| e | 0.980 | 1.058 (0.089) | 1.576 |
| f | 0.978 | 0.851 (0.149) | 1.065 |
| g | 0.974 | 1.087 (0.121) | 1.861 |

For each test case defined in Table 1 the parameters of the AEI model are fitted to the data extracted from the structures in the training set and used to calculate interaction energies for the structures in the test set. A sphere radius $r_{\text{sphere}} = 8.5$ Å is used. Correlation, slope, and RMSD with respect to fdP data are averaged over the conformations in the test set. The unsigned deviations of the slopes from 1, averaged over the conformations in the test set, are shown in parentheses. All atoms are assigned unit charges and 1-2 and 1-3 pairs are excluded. The unit of the RMSD is kcal/mol.

$u \in \{a, b, \ldots, g\}$. Excluding the training set $u = f$, the maximal deviation between any two $g_k^{(u)}$ is 0.0177 for $k = 1$, 0.0287 for $k = 2$, and 0.0229 for $k = 3$. This implies that these curves are very close to each other and basically overlap (see Fig. 2). Only $g_k^{(f)}$ differs significantly from the other curves. The maximal deviation between $g_k^{(f)}$ and all other fits for $k$ values of 1, 2, and 3 is 0.0367, 0.0774, and 0.0750, respectively. However, $g_k^{(f)}$ is expected to be an outlier: in small structures, most of the atoms are exposed to the solvent so that nearly all interactions experience large screening, that is, the average screening is higher than for a training set, which also includes structures with a large hydrophobic core. Therefore, for a given value of $u_{ij}^{\text{AEI}}/r_{ij}$, the average curve is biased towards high effective dielectric constants.

For each test case listed in Table 1 all pair interaction energies for every structure in the test set are calculated for unit charges, using the $g_k^{(u)}$ obtained by the fit based on the corresponding training set. The correlation, slope, and RMSD with respect to fdP data are determined for each structure in every test set. The results are summarized in Table 2. Note that only the results for $k = 3$ are shown, that is, 1-2 and 1-3 pairs are excluded. Taking all pairs into consideration merely improves results and is a less stringent test. None of the correlations is below 0.97 and apart from case (f), all slopes are close to 1 with the overall tendency being to overestimate rather than to underestimate energies. Applying $g_k^{(f)}$ (fit on proteins with less than 70 amino acids) to proteins with more than 70 amino acids [test set (f)] gives a slope of 0.85. The fit $g_k^{(f)}$ underestimates energies for large structures as it overestimates the average screening. The reverse effect, albeit less significant, can be observed for $g_k^{(g)}$. The fit on the training set consisting of proteins with more than 70 amino acids [training set (g)] slightly underestimates screening on average for the small structures. Because the structures in training set (g) include both buried and exposed atoms, the effect is hardly perceivable.

It is clear from the fits shown in Figure 2 and the cross correlation data given in Table 2 that far less than all the 52 structures are sufficient for the parameter optimization to converge [see test cases (a), (b), and (c)]. Furthermore, fitting on only folded, unfolded, or large structures does not introduce any bias [see test cases (d), (e), and (g)]. The model is highly independent of the shape of the proteins in the training set. However, a training set consisting of only small structures is not appropriate as it overestimates screening on average [see test case (f)].

The above analysis was carried out for 17 different values of the sphere radius $r_{\text{sphere}}$, namely $r_{\text{sphere}} = 6.0$ Å up to $r_{\text{sphere}} = 14.0$ Å with a step size of 0.5 Å. A value of $r_{\text{sphere}} = 8.5$ Å was found to perform best, but the model does not depend strongly on the radius. The results with a radius in the range from 7.5 to 9.0 Å differ only slightly. Clearly, a too small or too large sphere radius no longer discriminates whether an atom pair is in the bulk or on the surface, but there seems to be a relatively large range where this information is captured satisfactorily. Moreover, several combinations of different definitions of $\Theta_{ik}$ and different exponents for $u_{ij}^{\text{AEI}}$ [i.e., $(u_{ij}^{\text{AEI}})^{a/2}$ and $a$ was varied from 1.0 to 2.0 with a step size of 0.25] were investigated. Indeed, there are combinations that perform slightly better than the option presented here, but the marginal gain in accuracy is not considered to be worth the additional complexity in the formulas. In addition, a different weighting function is no longer intrinsic to the calculations in MD simulations. For all the following calculations, the functions $g_k$ derived from the fit on all conformations [training set (a)] with a sphere radius $r_{\text{sphere}} = 8.5$ Å will be used.

### *Comparison Between the AEI, GB, and SGM Models*

In the point charge approximation the total electrostatic interaction energy of a macromolecule in aqueous solution is

$$E_{\text{elec}}^{\text{inter}} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} E_{ij} \tag{12}$$

where $N$ is the number of charges in the system and $E_{ij}$ the screened Coulomb interaction energy of the $(i, j)$ pair. In the following sections, energies calculated by the AEI, GB, and SGM models are compared with the appropriate fdP values. In the next section, pair interaction energies $E_{ij}$ are analyzed, while in the following sections the sum of all interaction energies of atom $i$, $\sum_{j \neq i} E_{ij}$, and the total electrostatic interaction energy of the macromolecule, $E_{\text{elec}}^{\text{inter}}$, are discussed. The sums are useful to investigate cancellation of errors.

#### *Pair Interaction Energies $E_{ij}$*

Pair interaction energies $E_{ij}$ are calculated by the AEI, GB, and SGM models for each of the 52 conformations. The correlation, slope, and RMSD with respect to fdP data are evaluated and the results are shown in Figure 3 and Table 3. Unit charges are assigned to all atoms, and two sets of pairs are distinguished: all pairs and all but 1-2 and 1-3 pairs. Both the AEI and GB models perform distinctly better than the SGM model. The AEI and GB models show similar accuracy, and for most of the conformations the GB model has only a marginal advantage. However, the range in the correlation, slope, and RMSD are larger for the AEI than for

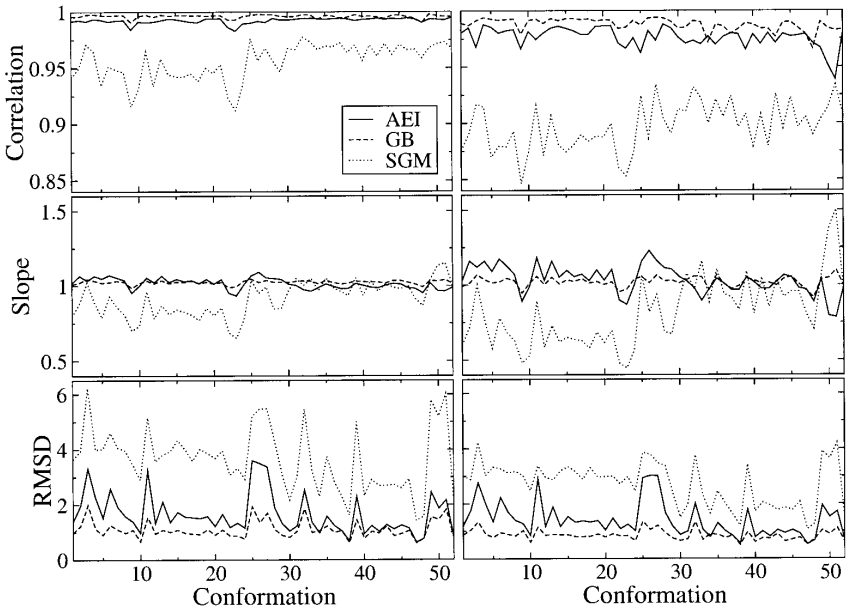


**Figure 3.** Pair interaction energies $E_{ij}$, calculated by the AEI [fit on training set (a), $r_{sphere}$ = 8.5 Å], GB, and SGM models, are compared to the corresponding fdP values. Correlation, slope, and RMSD are shown for each of the 52 conformations. Unit charges are assigned to all atoms and energy values are in kcal/mol. The data presented on the left hand column include all pairs and the data on the right hand column all but 1-2 and 1-3 pairs. The conformations are ordered such that the folded ones (conformations 1 to 29) precede the unfolded ones (conformations 30 to 52).

the GB model. In particular, the slopes vary more (see Table 3), but most of the single structure values are close to the mean value and the extremes are gathered in few conformations (see Fig. 3). From the data presented in Figure 3 one can deduce that both the AEI and GB models tend to overestimate interaction energies, that is, underestimate the screening effect.

A closer inspection of the data and structures reveals that conformations where most residues are exposed are less accurately represented in the AEI model compared to its average performance. Figure 4 shows pair interaction energies $E_{ij}$ as calculated by the AEI model against fdP data for a folded protein (2ins) and an unfolded structure [originating from a helix cut out from protein G (1pgb), named hlx1 in this article] that has the poorest correlation in the set of the 52 conformations. Note that the results for most of the structures are similar to 2ins. A strongly extended conformation of 17 residues (that is not in the set of the 52 conformations) gives a

**Table 3.** Minimal, Maximal, and Average Values of the Data Shown in Figure 3.

| | Correlation | | | Slope | | | RMSD | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Min | Max | Ave | Min | Max | Ave | Min | Max | Ave |
| All pairs | | | | | | | | | |
| AEI | 0.983 | 0.995 | 0.992 | 0.933 | 1.092 | 0.032 | 0.626 | 3.603 | 1.649 |
| GB | 0.990 | 0.998 | 0.996 | 0.986 | 1.045 | 0.025 | 0.603 | 1.992 | 1.102 |
| SGM | 0.912 | 0.977 | 0.956 | 0.651 | 1.152 | 0.113 | 1.410 | 6.230 | 3.661 |
| All but 1-2 and 1-3 pairs | | | | | | | | | |
| AEI | 0.939 | 0.988 | 0.976 | 0.784 | 1.228 | 0.081 | 0.513 | 3.019 | 1.400 |
| GB | 0.966 | 0.994 | 0.988 | 0.915 | 1.100 | 0.032 | 0.492 | 1.349 | 0.854 |
| SGM | 0.846 | 0.934 | 0.897 | 0.444 | 1.502 | 0.240 | 1.068 | 4.230 | 2.712 |

The average values shown in the slope column are the averages of the unsigned deviations of the slopes from 1. For the AEI model the fit on training set (a) with $r_{sphere}$ = 8.5 Å is used. The unit of the RMSD is kcal/mol.

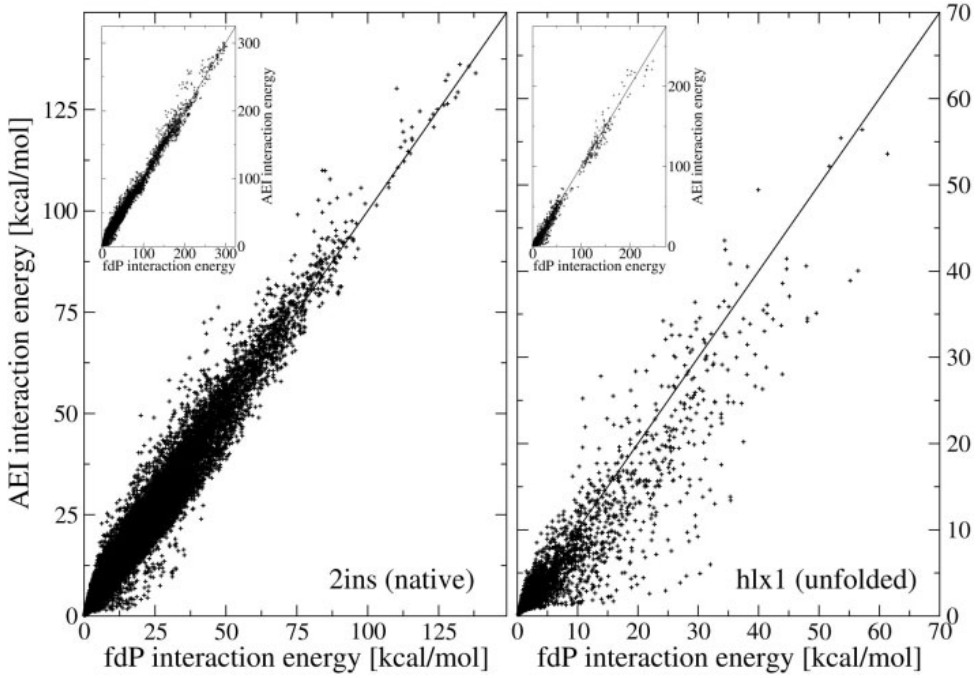


**Figure 4.** Pair interaction energies $E_{ij}$ for unit charges, calculated by the AEI model [fit on training set (a), $r_{\text{sphere}} = 8.5$ Å], are plotted against fdP values for all but 1-2 and 1-3 pairs for 2ins folded (left) and hlx1 unfolded (right). hlx1 unfolded is the conformation with the poorest correlation in the set of the 52 conformations. The insets show the interaction energies for all pairs. The unit of energy is kcal/mol.

correlation and slope of only 0.91 and 0.66, respectively (excluding 1-2 and 1-3 pairs). The AEI model underestimates interaction energies for very extended structures. However, this is not a serious disadvantage because they are not realistic and are not usually sampled in conventional MD simulations.

The measure of enclosure $u_{ij}^{\text{AEI}}$ presented in this work is applicable for any macromolecular system. Yet, for peptides (20 residues or less) results can be improved by a slightly different choice. Using $(u_{ij}^{\text{AEI}})^{1/2}$ instead of $u_{ij}^{\text{AEI}}$ improves the accuracy for very extended conformations without any major deterioration of the values of the folded peptides.

To further compare the accuracy of the three models, Figure 5A shows a histogram of the deviations of the interaction energies calculated by the AEI, GB, and SGM models from the fdP values. Again, the GB model is slightly more accurate than the AEI model, whereas the SGM model has by far the largest errors. The deviations are essentially the same if 1-2 and 1-3 pairs are included in the calculations.

### *Error Cancellation for Atomic Energies*

The relevant quantities in MD simulations are sums over pair interaction energies and their derivatives and not single pair values. The model that best reproduces $E_{ij}$ with respect to fdP data is not necessarily the best at reproducing $\Sigma_{j \neq i} E_{ij}$ if the errors in $E_{ij}$ cancel each other poorly. There is a fortuitous cancellation of errors in the GB model because a systematic overestimation (or underestimation) of the effective Born radii has a compensating effect on sums of interactions involving like and opposite charges.[29] In the following, the cancellation of errors in the AEI, GB, and SGM models is compared. For this purpose all atoms of the 52 conformations are assigned partial charges. Let $E_{ij}^{\text{fdP}}$ denote the interaction energy calculated by fdP and $E_{ij}$ the energy calculated by the AEI, GB, or SGM model. In a first step the error of each pair interaction energy $E_{ij}$ with respect to the fdP value, that is, $E_{ij} - E_{ij}^{\text{fdP}}$, is determined. Then, for each nonzero partial charge $i$ individually (a total of 37,869 charges), positive errors ($E_{ij} - E_{ij}^{\text{fdP}} > 0$) and negative errors ($E_{ij} - E_{ij}^{\text{fdP}} < 0$) are added up separately. The total error of charge $i$ is the sum of the two and the values are shown in Figure 6. The sums $\Sigma_{j \neq i} E_{ij}$, as calculated by the AEI, GB, and SGM models for each atom $i$ with a nonzero partial charge, are plotted against the appropriate fdP values in Figure 7. The corresponding correlation, slope, and RMSD are shown in Table 4. Note that the vertical deviations from the diagonal line in Figure 7 that are used to calculate the RMSD correspond to the total errors in Figure 6. The correlation and slope, however, cannot be deduced from the data in Figure 6. A histogram of the distri-

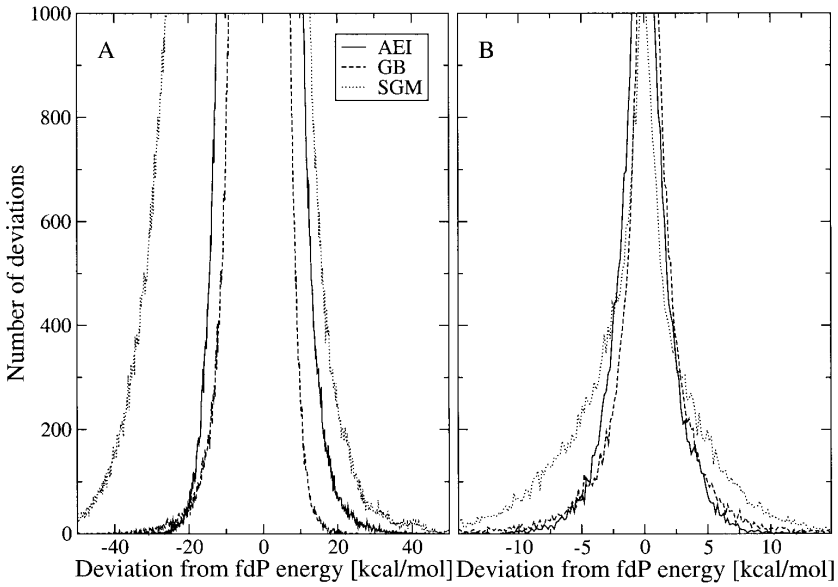


**Figure 5.** Distribution of deviations from fdP values of pair interaction energies $E_{ij}$ (A) and the sums $\Sigma_{j \neq i} E_{ij}$ (B), calculated by the AEI, GB, and SGM models. A total of 625 and 185 bins of 0.16 kcal/mol each span the energy intervals given on the abscissa axes in (A) and (B), respectively. The number of data points for which the deviation in energy falls within a given bin is shown on the ordinate axis. For (A), all atoms are assigned unit charges, whereas for (B), partial charges are used. In both cases 1-2 and 1-3 pairs are excluded from the calculations.

bution of errors is shown in Figure 5B. Moreover, Table 4 shows the results if the sums $\Sigma_{j \neq i} E_{ij}$ are calculated with a cutoff of 7.5 Å (the default cutoff of CHARMM PARAM19) in the AEI, GB, and SGM models and compared to the corresponding fdP values obtained by adding up interaction energies with no cutoff. Note that 1-2 and 1-3 pairs are excluded from the data presented in Figures 5B, 6, and 7 and Table 4, but the results look similar if all pairs are taken into account.

All three models do in fact benefit greatly from cancellation of errors. Clearly, the SGM model has the largest total errors. It is interesting to note that for small errors (<5 kcal/mol) the AEI and GB models show essentially the same frequencies, whereas larger errors (>5 kcal/mol) occur less often in the AEI than in the GB model (see Fig. 7 and tails in Fig. 5B). From the data presented in Figures 5B and 7 one can deduce that in the case of an infinite cutoff, errors cancel each other better in the AEI than in the GB model. For a cutoff of 7.5 Å the two models show similar accuracy (Table 4).

### *Total Electrostatic Interaction Energy of Native and Non-Native Conformations*

It is important to test the accuracy of the total electrostatic interaction energy calculated by the AEI model for different conformations of the same macromolecule. For this purpose high temperature unfolding simulations at 450 K for 20 ns using an implicit solvation model[27] of a SH3 domain (1shg, 57 residues) and a three-stranded antiparallel $\beta$-sheet (beta3s, 20 residues[30]) were performed. Coordinates were saved every 5 ps and all snapshots were sorted according to increasing radius of gyration (Rg). Then 100 conformations were chosen as follows: every 25th conformation of the 500 snapshots with the lowest Rg (20 conformations), every 25th conformation of the 500 snapshots with the largest Rg (20 conformations), and every 50th conformation of the remaining 3000 snapshots (60 conformations). Furthermore, the native state was added. The conformations ranged from folded to significantly extended. They covered a range in the radius of gyration from 10.2 to 25.4 Å for 1shg and from 6.9 to 12.3 Å for beta3s. Note that from the 101 conformations of 1shg or beta3s, only the native state and one of the unfolded states were in training set (a) that was used to parameterize the AEI model. The comparison was limited to two proteins because of the large computational requirements for the fdP calculations on the set of 100 conformations. For each structure the total electrostatic interaction energy $E_{\mathrm{elec}}^{\mathrm{inter}}$ [see eq. (12)] was calculated. The results for the AEI, GB, and SGM models are compared to the fdP data and are shown in Figure 8 and Table 5. Also shown in Table 5 are the results if $E_{\mathrm{elec}}^{\mathrm{inter}}$ is calculated in the AEI, GB, and SGM models with a cutoff of 7.5 Å and compared to the appropriate fdP data obtained with no cutoff. The plots in Figure 8 indicate that the AEI model is accurate enough not only for compact and unfolded structures but also for conformations

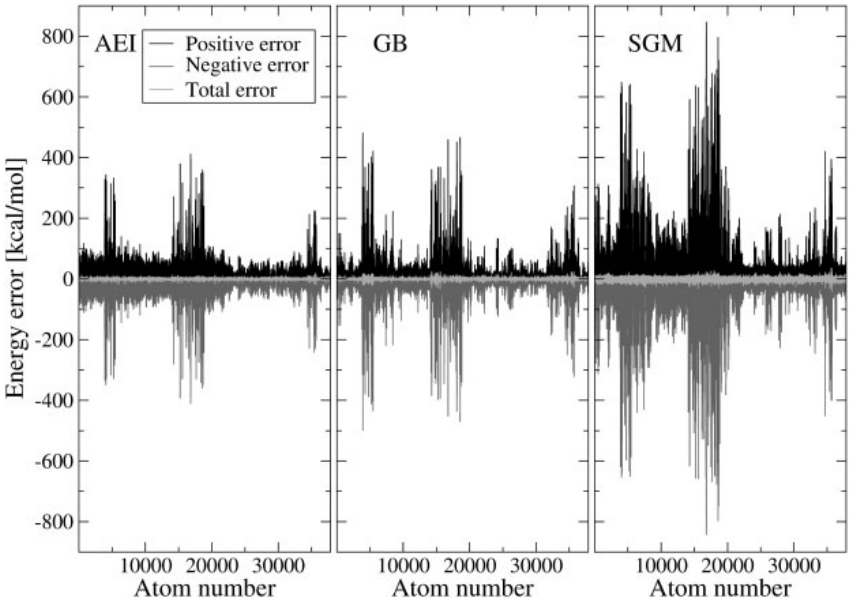


**Figure 6.** Error cancellation for pair interaction energies as calculated by the AEI [fit on training set (a), $r_{\text{sphere}} = 8.5$ Å], GB, and SGM models. For each atom $i$ with nonzero partial charge, the total error $\Sigma_{j \neq i} (E_{ij} - E_{ij}^{\text{fdP}})$ is given in light gray. The positive and negative contributions to the total error are shown in black and gray, respectively. Partial charges are used and 1-2 and 1-3 pairs are excluded.

with an intermediate value of the Rg. Furthermore, the RMSD values of the total electrostatic interaction energy are smaller in the AEI than in the GB approach. There is a systematic shift towards lower and higher energy values for the AEI and GB model, respectively (see Fig. 8), and the shift is smaller in the AEI than in the GB model. Note that the electrostatic interaction energy alone is not expected to discriminate the native state from other compact conformations.

### *Efficiency*

Finally, we comment on the computational requirements. The AEI model is highly efficient; it is only 10% slower than *vacuo,* irrespective of the size of the molecule. According to ref. 13, the GB approach is slower by a factor of 5 (for large proteins with more than 60 amino acids) to 10 (for small proteins and peptides with up to 60 amino acids) compared to *vacuo.*

### *Solvation Energies*

To further assess the physical relevance of a measure of enclosure based on the sum over weighted volumes of neighbors, this section addresses the evaluation of solvation energies in the framework of the AEI model. A brief description is given here because the focus of this article is on screened interaction energies.

In the Methods section a measure of enclosure for a pair of atoms $(i, j)$ is introduced. In the same spirit one can define a measure of enclosure for a single atom $i$ by

$$w_i^{\text{AEI}} = \sum_{k=1}^{N} v_k \Theta_{ik} \tag{13}$$

where $v_k$ is the van der Waals volume of atom $k$, $N$ the total number of atoms in the system, and $\Theta_{ik}$ is defined in the Methods section with $r_{\text{sphere}} = 8.5$ Å. Note that $\Theta_{ik}$ is different from zero only for the atoms in a sphere of radius 8.5 Å centered on atom $i$. Analytical functions of $w_i^{\text{AEI}}$, that is, $\Delta E_i^{\text{AEI}} = h_p(w_i^{\text{AEI}})$, where $\Delta E_i^{\text{AEI}}$ denotes the solvation energy of atom $i$ calculated in the AEI model, are fitted to fdP-derived solvation energies for unit charges of the atoms of one protein (1a2p, 1,073 atoms). The index $p$ accounts for the fact that different functions are necessary for different ranges of the van der Waals radii (see Appendix). The AEI solvation energies for all atoms with nonzero partial charge of 10 proteins not used to parameterize the model (1bpi, 1crn, 1hdn, 1pgb, 1pht, 1ycq, 1ycr, 2ci2, 2ptl, beta3s), are shown in Figure 9. For the AEI model, the correlation and slope are 0.987 and 0.952, respectively, and for the GB model 0.986 and 0.632, respectively. These results indicate that a measure of enclosure for a single atom $i$ based on summing over neighbors allows the modeling of not only interaction but also solvation energies.

## Conclusion

Both the AEI and GB models utilize a measure of enclosure for pairs of charges to calculate the screened electrostatic interac-

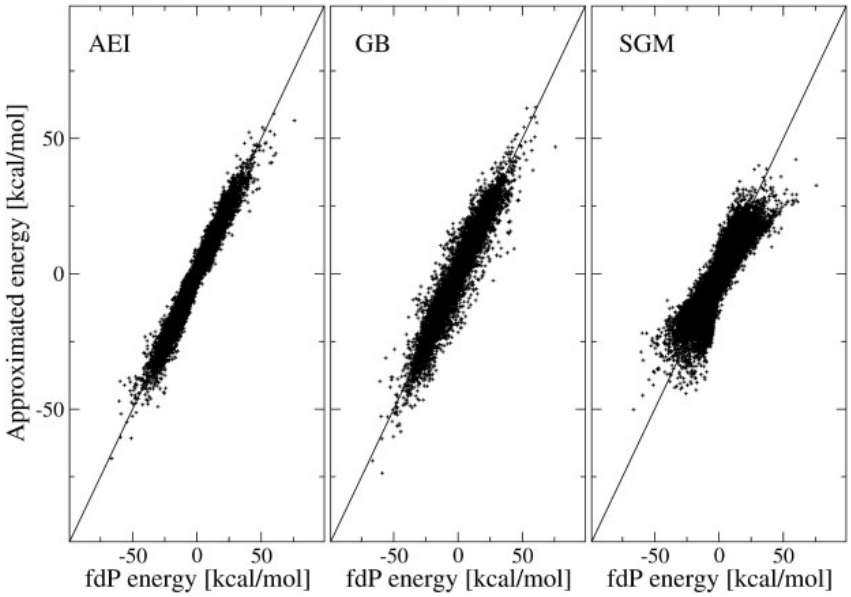


**Figure 7.** Each data point represents the sum of all pair interaction energies of an atom $i$ with nonzero partial charge, $\Sigma_{j \neq i} E_{ij}$. The AEI [fit on training set (a), $r_{\text{sphere}} = 8.5$ Å], GB, and SGM values are plotted against the fdP data. Partial charges are used and 1-2 and 1-3 pairs are excluded. See text for details.

tion energy. For each charge in the system an effective volume and an effective Born radius is evaluated in the AEI and the GB approach, respectively. These quantities are combined in a heuristic way to obtain a measure of enclosure for a pair of charges. The essential element of the AEI model is to define such a measure with information easily available from a reasonably large neighborhood of a given pair. The degree of enclosure of two charges quantifies the distribution of solute atomic volumes surrounding the pair. The appealing feature of the AEI model is the efficiency with which the measure of enclosure can be calculated. It is simply the square of the sum of weighted atomic volumes. Hence, the present implementation of the AEI model uses only quantities whose calculation is intrinsic to MD simulations so that the computational overhead is negligible with respect to *vacuo.* In the GB approach the measure of enclosure of a pair of charges is the square root of the product of their effective Born radii, whose calculation requires integration of the electrostatic energy density over the solute volume. The sum of all pair interaction energies of an atom $i$ (relevant for MD simulations) and the total electrostatic interaction energy of a macromolecule are reproduced more accurately in the AEI than in the GB approach. Only single pair interaction energies are slightly better approximated in the GB model. The validity of the AEI model is further assessed by

**Table 4.** Comparison of the Sum of All Interaction Energies of Atom $i$, $\Sigma_{j \neq i} E_{ij}$, Calculated by the AEI, GB, and SGM Models and fdP.

| | No cutoff | | | Cutoff of 7.5 Å | | |
|---|---|---|---|---|---|---|
| Model | Correlation | Slope | RMSD | Correlation | Slope | RMSD |
| AEI | 0.981 | 1.016 | 2.152 | 0.957 | 0.994 | 3.258 |
| GB | 0.966 | 1.003 | 2.895 | 0.955 | 0.989 | 3.312 |
| SGM | 0.890 | 0.794 | 4.958 | 0.878 | 0.782 | 5.197 |

For the AEI model the fit on training set (a) with $r_{\text{sphere}} = 8.5$ Å is used. Partial charges are assigned to all atoms and 1-2 and 1-3 pairs are excluded. No cutoff is applied for the calculation of the data on the left hand part of the table. For the data on the right hand part, the sums $\Sigma_{j \neq i} E_{ij}$ are calculated with a cutoff of 7.5 Å in the AEI, GB, and SGM models and compared to the corresponding fdP values obtained with no cutoff. The unit of the RMSD is kcal/mol.

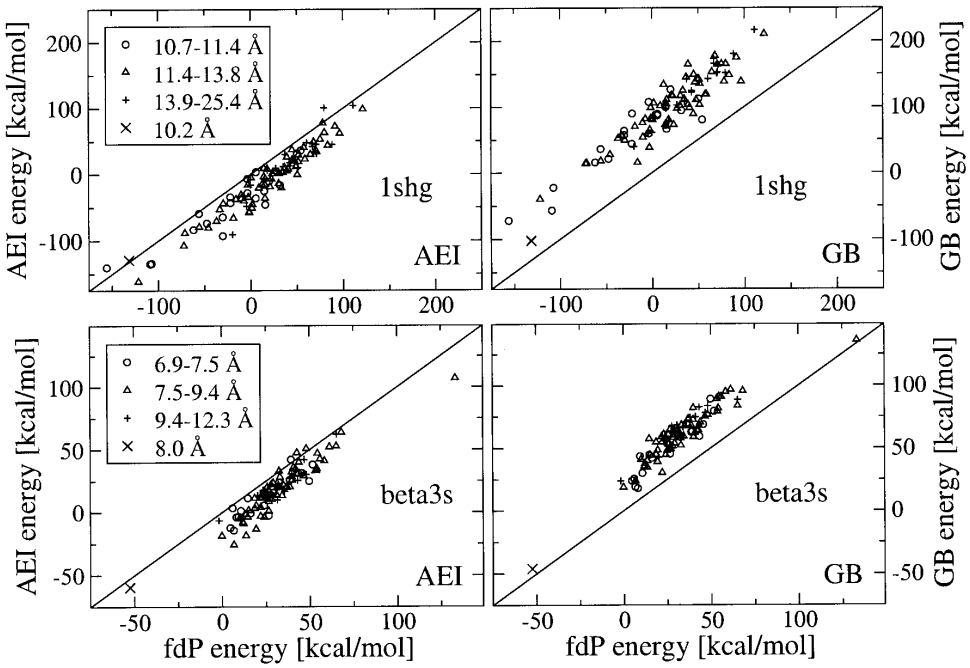


**Figure 8.** Each data point represents the total electrostatic interaction energy $E_{\text{elec}}^{\text{inter}}$ for a given conformation as calculated by the AEI [fit on training set (a), $r_{\text{sphere}}$ = 8.5 Å] and GB models against the corresponding fdP value. Data for 101 conformations along a high temperature unfolding trajectory of 1shg (top) and beta3s (bottom) are shown. Partial charges are used and 1-2 and 1-3 pairs are excluded from these calculations. Different symbols discriminate between different ranges for the radius of gyration. Circles and pluses represent the 20 conformations with small and large Rg, respectively, and triangles the 60 intermediate ones. The total electrostatic interaction energy of the native state is shown by the symbol x.

**Table 5.** For 101 Conformations along a High Temperature Unfolding Trajectory of 1shg and beta3s, the Total Electrostatic Interaction Energy $E_{\text{elec}}^{\text{inter}}$ Is Calculated by the AEI, GB, and SGM Models.

| | No cutoff | | | Cutoff of 7.5 Å | | |
|---|---|---|---|---|---|---|
| Model | Correlation | Slope | RMSD | Correlation | Slope | RMSD |
| 1shg | | | | | | |
| AEI | 0.955 | 0.983 | 31.265 | 0.917 | 0.969 | 27.631 |
| GB | 0.941 | 1.040 | 81.515 | 0.935 | 1.038 | 83.598 |
| SGM | 0.706 | 0.674 | 139.198 | 0.661 | 0.672 | 124.610 |
| beta3s | | | | | | |
| AEI | 0.939 | 1.006 | 15.159 | 0.916 | 0.958 | 15.567 |
| GB | 0.937 | 1.043 | 28.923 | 0.925 | 1.006 | 27.835 |
| SGM | 0.818 | 0.849 | 63.504 | 0.769 | 0.854 | 63.163 |

For the AEI model, the fit on training set (a), $r_{\text{sphere}}$ = 8.5 Å is used. Correlation, slope, and RMSD with respect to fdP data are shown. Partial charges are used and 1-2 and 1-3 pairs are excluded from these calculations. No cutoff is applied for the calculation of the data on the left hand part of the table. The data on the right hand part show the total electrostatic interaction energy calculated with a cutoff of 7.5 Å in the AEI, GB, and SGM models, compared to the corresponding fdP values obtained with no cutoff. The unit of the RMSD is kcal/mol.

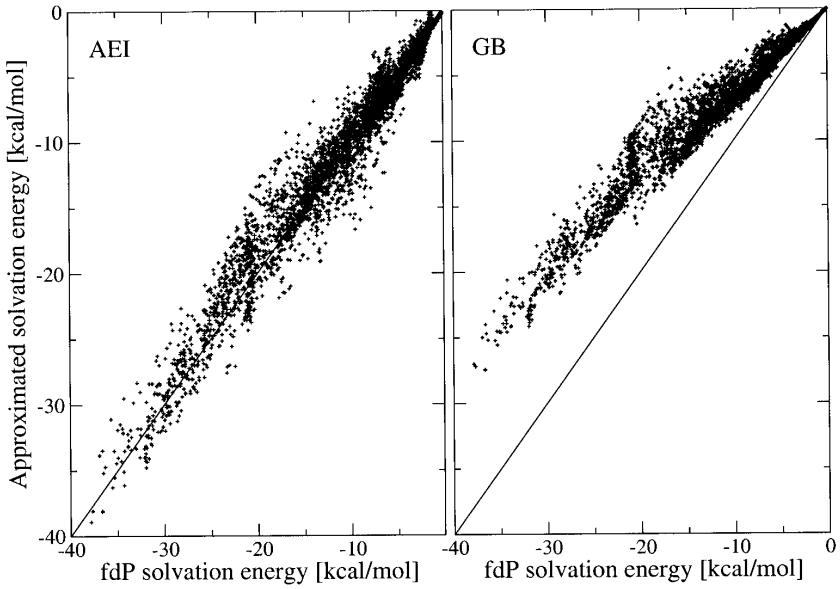


**Figure 9.** Atomic solvation energies calculated by the AEI (left) and GB (right) models for 10 proteins (6504 data points) are compared with the fdP values. Partial charges are assigned to all atoms.

demonstrating that solvation energies can be calculated with a measure of enclosure for single atoms that is similar to the one used for pairs.

## Acknowledgments

We thank two anonymous referees for very useful comments.

## Appendix

### *Interaction Energy*

Let $r_k$ be the van der Waals radius of atom $k$ and $v_k = \frac{4}{3}\pi r_k^3$ its van der Waals volume. The van der Waals radii are taken from the CHARMM parameter set PARAM19 and depend only on the atom type. The measure of enclosure $u_{ij}^{\mathrm{AEI}}$ is defined by eqs. (5) and (7). Let $r_{ij}$ be the distance between atoms $i$ and $j$ and define $p_{ij} = u_{ij}^{\mathrm{AEI}}/r_{ij}$. The reciprocal of the effective dielectric function in the AEI model is denoted by $g_k$, where $k = 1$ for 1-2 pairs, $k = 2$ for 1-3 pairs, and $k = 3$ for all but 1-2 and 1-3 pairs. Due to the close relationship between the AEI and GB approaches, the $g_k$ are chosen to be of the same functional form as the reciprocal of the effective dielectric function in the GB model [see $f$ in eq. (4)]:

$$g_k(p_{ij}) = a_{1,k} - (a_{1,k} - a_{2,k}) \times (1 + (a_{4,k}(p_{ij} + a_{3,k}))^2 e^{-[1/(a_{5,k}(p_{ij}+a_{3,k}))^2]})^{-(1/2)} \quad \text{(A1)}$$

All five parameters $a_{i,k}$ appearing in $g_k$ have a well defined meaning. The functions $g_k$ are of sigmoidal shape with the maximum and minimum value $a_{1,k}$ and $a_{2,k}$, respectively, if $a_{1,k} > a_{2,k}$. This condition is always satisfied in the AEI model. The parameter $a_{3,k}$ translates the function $g_k$ parallel to the abscissa axis, and $a_{4,k}$ and $a_{5,k}$ are scaling factors. The parameters $a_{i,k}$ are determined by fitting $g_k$ to the inverse of fdP-derived effective dielectric constants extracted from all 52 conformations [training set (a)]. The sphere radius used is $r_{\mathrm{sphere}} = 8.5$ Å. The parameters are determined to be

$$(a_{i,k}) = \begin{pmatrix} +0.113 \cdot 10^{+01} & +0.133 \cdot 10^{+01} & +0.100 \cdot 10^{+01} \\ +0.253 \cdot 10^{+00} & -0.980 \cdot 10^{-01} & +0.127 \cdot 10^{-01} \\ +0.145 \cdot 10^{+07} & +0.124 \cdot 10^{+07} & +0.000 \cdot 10^{+00} \\ +0.451 \cdot 10^{-06} & +0.541 \cdot 10^{-06} & +0.135 \cdot 10^{-05} \\ +0.998 \cdot 10^{+00} & +0.967 \cdot 10^{+00} & +0.581 \cdot 10^{-05} \end{pmatrix} \quad \text{(A2)}$$

The functions $g_1$ and $g_2$ have five parameters each whereas the function $g_3$, which is the most relevant for molecular mechanics and dynamics, has in effect only two parameters because $a_{1,3}$, $a_{2,3}$, and $a_{3,3}$ are set to the standard GB values, that is, $a_{1,3} = 1/\varepsilon_m = 1$, $a_{2,3} = 1/\varepsilon_s = 1/78.5$, and $a_{3,3} = 0$.

### *Solvation Energy*

The measure of enclosure $w_i^{\mathrm{AEI}}$ for a single atom $i$ is defined by eq. (13). Solvation energies are calculated by $h_p(w_i^{\mathrm{AEI}}) = b_{1,p} + b_{2,p} w_i^{\mathrm{AEI}} + b_{3,p}(w_i^{\mathrm{AEI}})^2$, where $p = 1$ for van der Waals radii in

the range from 0.5–1.0 Å, $p = 2$ for the range from 1.5–2.0 Å (there are no van der Waals radii in PARAM19 with a value between 1.0 and 1.5 Å), and $p = 3$ for van der Waals radii larger than 2.0 Å. The sphere radius used is $r_{\text{sphere}} = 8.5$ Å. The parameters $b_{i,p}$ are determined by fitting the functions $h_p$ to fdP-derived atomic solvation energies for unit charges of one single protein (1a2p, 1,073 atoms). The parameters are determined to be

$$(b_{i,p}) = \begin{pmatrix} -0.220 \cdot 10^{+03} & -0.108 \cdot 10^{+03} & -0.766 \cdot 10^{+02} \\ +0.209 \cdot 10^{+00} & +0.502 \cdot 10^{-01} & +0.169 \cdot 10^{-01} \\ -0.326 \cdot 10^{-04} & +0.193 \cdot 10^{-04} & +0.250 \cdot 10^{-04} \end{pmatrix} \tag{A3}$$

## References


1. Gilson, M. K. Curr Opin Struct Biol 1995, 5, 216.
2. Roux, B.; Simonson, T. Biophys Chem 1999, 78, 1.
3. Tomasi, J.; Persico, M. Chem Rev 1994, 94, 2027.
4. Cramer, C. J.; Trulhar, D. G. Chem Rev 1999, 99, 2161.
5. Orozco, M.; Luque, F. J. Chem Rev 2000, 100, 4187.
6. Warwicker, J.; Watson, H. C. J Mol Biol 1982, 157, 671.
7. Gilson, M. K.; Honig, B. H. Proteins Struct Funct Genet 1988, 4, 7.
8. Bashford, D.; Karplus, M. Biochemistry 1990, 29, 10219.
9. Davis, M. E.; Madura, J. D.; Luty, B. A.; McCammon, J. A. Comput Phys Commun 1991, 62, 187.
10. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. J Am Chem Soc 1990, 112, 6127.
11. Scarsi, M.; Apostolakis, J.; Caflisch, A. J Phys Chem A 1997, 101, 8098.
12. Bashford, D.; Case, D. A. Annu Rev Phys Chem 2000, 51, 129.
13. Lee, M. S.; Salsbury Jr, F. R.; Brooks III, C. L. J Chem Phys 2002, 116, 10606.
14. Hawkins, G. D.; Cramer, C. J.; Trulhar, D. G. Chem Phys Lett 1995, 246, 122.
15. Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. J Phys Chem A 1997, 101, 3005.
16. Dominy, B. N.; Brooks III, C. L. J Phys Chem B 1999, 103, 3765.
17. Warshel, A.; Levitt, M. J Mol Biol 1976, 103, 227.
18. Gelin, B. R.; Karplus, M. Biochemistry 1979, 18, 1256.
19. Mehler, E. L.; Eichele, G. Biochemistry 1984, 23, 3887.
20. Mehler, E. L. Protein Eng 1990, 3, 415.
21. Hassan, S. A.; Guarnieri, F.; Mehler, E. L. J Phys Chem B 2000, 104, 6478.
22. Wang, L.; Hingerty, B. E.; Srinivasan, A. R.; Olson, W. K.; Broyde, S. Biophys J 2000, 83, 382.
23. Mallik, B.; Masunov, A.; Lazaridis, T. J Comput Chem 2002, 23, 1090.
24. Hirschfelder, J. O.; Curtiss, C. F.; Bird, R. B. Molecular theory of gases and liquids; John Wiley & Sons: New York, 1964.
25. Jackson, J. D. Classical Electrodynamics; John Wiley & Sons: New York, 1975.
26. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187.
27. Ferrara, P.; Apostolakis, J.; Caflisch, A. Proteins Struct Funct Genet 2002, 46, 24.
28. Im, W.; Beglov, D.; Roux, B. Comput Phys Commun 1998, 111, 59.
29. Onufriev, A.; Case, D. A.; Bashford, D. J Comput Chem 2002, 23, 1297.
30. De Alba, E.; Santoro, J.; Rico, M.; Jimenez, M. A. Protein Sci 1999, 8, 854.

# Chapter 5

# FACTS: Fast Analytical Continuum Treatment of Solvation

# FACTS: Fast Analytical Continuum Treatment of Solvation

Urs Haberthür and Amedeo Caflisch*

Department of Biochemistry, University of Zurich

Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

E-mail: caflisch@bioc.unizh.ch

Phone: +41 44 635 55 21

Fax: +41 44 635 68 62

October 16, 2006

**Abstract**

An efficient method for calculating the free energy of solvation of a (macro)molecule embedded in a continuum solvent is presented. It is based on the fully analytical evaluation of the volume and spatial symmetry of the solvent that is displaced from around a solute atom by its neighboring atoms. The two measures of solvent displacement are combined in empirical equations to approximate the atomic (or self) electrostatic solvation energy and the solvent accessible surface area. The former directly yields the effective Born radius, which is used in the generalized Born (GB) formula to calculate the solvent-screened electrostatic interaction energy. A comparison with finite difference Poisson data shows that atomic solvation energies, pair interaction energies and their sums are evaluated with a precision comparable to the most accurate GB implementations. Furthermore, solvation energies of a large set of protein conformations have an error of only 1.5%. The solvent accessible surface area is used to approximate the non-polar contribution to solvation. The approach, called FACTS (Fast Analytical Continuum Treatment of Solvation), is only four times slower than using the vacuum energy in molecular dynamics (MD) simulations of proteins. Notably, the folded state of structured peptides and proteins is stable at room temperature in 100-ns MD simulations using FACTS and the CHARMM force field.

*Corresponding author

# 1 Introduction

An accurate treatment of the effects of aqueous solvent in molecular dynamics (MD) simulations of biological (macro)molecules is of key importance because cells and physiological fluids consist mainly of water. The exact calculation of the electrostatic energy of a protein in solution requires the evaluation of the interactions among all solute-solute, solute-solvent, and solvent-solvent pairs of charges. However, this is computationally expensive for fully hydrated macromolecules. Despite continuous advances in both the development of parallel MD simulation code and the performance of ordinary low cost computer processors, explicit solvent MD simulations of large proteins lasting longer than 100 nanoseconds are still almost prohibitive. A simplified treatment that does not require the solvent degrees of freedom and interaction centers explicitly can be very useful, and for large systems it represents the only affordable description of the solvent.

The essential concept in continuum electrostatics models is to represent the solvent as a featureless high dielectric medium, and to describe the macromolecule as a region with a low dielectric constant and a spatial charge distribution [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. In this way, the solvent degrees of freedom and interaction centers are not taken into account explicitly. The Poisson equation provides an exact description of such a solute/solvent system. The numerical solution of the finite difference Poisson (fdP) equation [12, 13, 14, 15] is more efficient than the explicit treatment of the solvent but still not fast enough for effective utilization in computer simulations of macromolecules.

The generalized Born (GB) model was introduced to facilitate an efficient evaluation of continuum electrostatic energies [16]. The most critical aspect of the GB model is the calculation of the *effective Born radii* which measure the degree of burial of individual solute charges. This measure is combined in a heuristic way to obtain a correction to the Coulomb law for each atom pair [16]. In contrast to implicit solvation models which use a distance-dependent dielectric function [17, 18, 19], the GB equation takes into account the effect of both the charge-charge distance and the degree of solvent exposure of the interacting charges. Accurate GB implementations published as of today are between 20 and 40 times slower than simulations in vacuo [20]. Moreover, for proteins of about 100 residues the computational cost per MD time step is about the same for accurate GB models and explicit water simulations with periodic boundary conditions [21].

Water molecules in the liquid state influence the electrostatic energy of a macromolecule in two ways. They solvate each individual charge of the solute (*atomic solvation energy*), and they screen the interaction between charge pairs [22]. In a previous work, we introduced a geometric measure of the degree of burial of *pairs* of interacting solute charges for quantifying the screening effect [23]. The aim of the present article is to adopt similar steric concepts for the efficient evaluation of the effective Born radius (which is inversely proportional to the atomic solvation energy) using the local environment of each solute atom. The same geometric formalism is also proposed for the calculation of the solvent accessible surface area (SASA) of individual atoms of the solute, which is used for approximating the nonpolar contribution to solvation. The resulting continuum model, called FACTS, is a fully

analytical and comprehensive treatment of solvation effects. A comparison is given with one of the most accurate GB methods [20], i.e., GB using molecular volume (GBMV [21]). The extensive validation provides evidence that FACTS is as accurate as the best available GB implementations, and MD simulations with FACTS are only four times slower than using the energy in vacuo.

# 2 Methods

## 2.1 Continuum Electrostatics

The electrostatic potential $\phi$ of a charge distribution $\rho$, given the dielectric function $\varepsilon$, is uniquely defined by the Poisson equation $\nabla\varepsilon\nabla\phi = -4\pi\rho$ and appropriate boundary conditions. Consider a macromolecule immersed in a solvent. For an explicit treatment of both solute and solvent atoms $\varepsilon = 1$ everywhere, and the Poisson equation reduces to $\Delta\phi = -4\pi\rho$. If additionally the charge density $\rho$ is a superposition of spherically symmetric charge distributions within hard (impenetrable) spheres, the Coulomb potential is recovered from the Poisson equation. While such an explicit treatment is mathematically simple and accurate in the framework of classical physics, the computational costs for (macro)molecular simulations can be very high. For instance, sampling a statistically significant number of folding and unfolding transitions of structured peptides at equilibrium require simulations in the 1-10 microseconds timescale [24].

The most efficient and widely used approximation of electrostatic solvation effects is the continuum model [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. The volume occupied by the solute (macromolecule) is assigned a low dielectric constant $\varepsilon_m$ (typically 1, 2, or 4) and the charge distribution is defined by the partial charges of the macromolecular atoms. The solvent is replaced by a uniform medium with a high dielectric constant $\varepsilon_s$ (typically 78.5 or 80 in the case of water). The boundary between low and high dielectric regions is usually defined by the molecular surface which is spanned by the *surface* of a solvent probe sphere rolled over the van der Waals envelope of the solute. Compared to the van der Waals surface the molecular surface avoids inner cavities, and compared to the solvent accessible surface it yields hydration free energies that are in better agreement with experimental data [25].

## 2.2 Generalized Born (GB) Model

The GB approach [16] is an efficient analytical approximation to the solution of the Poisson equation (see also the review articles [5, 8]). Its derivation starts from a simple system (i.e., very large interatomic distances), for which an analytical solution exists, and proceeds by extending this solution in a heuristic way so that it becomes applicable to any macromolecular configuration. Consider a system consisting of $N$ atoms with charges $q_i$ subject to the following two conditions. Firstly, each atom is treated as a hard sphere with a spherically symmetric charge distribution that is located on its surface. Secondly, the atoms are separated by large distances, i.e., $r_i^{vdW} \ll r_{ij}$ where $r_{ij} = |\vec{x}_i - \vec{x}_j|$ and $r_i^{vdW}$ is the van der Waals radius of atom $i$. For this system, the electrostatic contribution to the solvation free

energy yields [22]

$$\Delta G^{el,\infty} = -\frac{\tau}{2}\sum_{i=1}^{N}\frac{q_i^2}{r_i^{vdW}} - \tau\sum_{1\le i<j\le N}\frac{q_iq_j}{r_{ij}} \quad (1)$$

$$= -\frac{\tau}{2}\sum_{i,j=1}^{N}\frac{q_iq_j}{r_{ij}} \quad (2)$$

where $\tau = \frac{1}{\varepsilon_m} - \frac{1}{\varepsilon_s}$ and $r_{ii} = r_i^{vdW}$ in equation (2). (A multiplicative factor of 332.0716 is used in front of $\tau$ for obtaining energy values in kcal/mol with interatomic distances in Å and partial charges in electronic units). If the atoms approach each other, i.e., $r_i^{vdW} \approx r_{ij}$, their spatial extent can no longer be neglected and the solution of the Poisson equation is no longer given by the superposition of electrostatic fields generated by point charges. The aim of the GB model is to modify equation (2) such as to make it accurate also for configurations where atoms are close to each other as in a macromolecule. The idea is to replace $r_{ij}$ by an effective interatomic distance $r_{ij}^{eff}$ to approximate the screening effect. For this purpose, the effective Born radius $R_i$ of atom $i$ is defined by

$$R_i = -\frac{\tau q_i^2}{2\Delta G_i^{el}} \quad (3)$$

where $\Delta G_i^{el}$ denotes the electrostatic solvation free energy of atom $i$, which is the solvation free energy of the macromolecule when all charges, except the one of atom $i$, are set to zero. For a hypothetical spherical macromolecule with atom $i$ at its center, $\Delta G_i^{el}$ is the solvation energy of a single ion with charge $q_i$ and a van der Waals radius equal to the radius of the macromolecule [26]. In this case the effective Born radius $R_i$ is simply the macromolecular radius. If the macromolecule is not spherical or atom $i$ does not sit at the center of the sphere, the effective Born radius of atom $i$ approximates its average distance from the surface of the macromolecule [27]. Equation (2) is modified by performing the substitution [16]

$$r_{ij} \mapsto r_{ij}^{eff} = \sqrt{r_{ij}^2 + R_iR_j\exp(-r_{ij}^2/\kappa R_iR_j)} \quad (4)$$

which yields

$$\Delta G^{el,GB} = \sum_{i=1}^{N}\Delta G_i^{el} - \tau\sum_{1\le i<j\le N}\frac{q_iq_j}{\sqrt{r_{ij}^2 + R_iR_j\exp(-r_{ij}^2/\kappa R_iR_j)}} \quad (5)$$

$$= -\frac{\tau}{2}\sum_{i,j=1}^{N}\frac{q_iq_j}{\sqrt{r_{ij}^2 + R_iR_j\exp(-r_{ij}^2/\kappa R_iR_j)}} \quad (6)$$

where the $r_{ii}$ are set to zero and $\kappa$ is a constant usually set to 4 or 8. Note that in the limit $r_{ij} \to \infty$ equation (6) reduces to equation (2), i.e., $\lim_{r_{ij}\to\infty}\Delta G^{el,GB} = \Delta G^{el,\infty}$ since $\lim_{r_{ij}\to\infty} r_{ij}^{eff} = r_{ij}$ and $\lim_{r_{ij}\to\infty} R_i = r_i^{vdW}$. Equation (6) is only semi-analytical as it requires the evaluation of the effective Born radii. In fact, at this stage the GB

approach has merely shifted the calculation of $\Delta G^{el}$ for the macromolecule to the evaluation of $\Delta G_i^{el}$ for each atom. An important observation is that equation (6) yields very accurate results if $\Delta G_i^{el}$ is a good approximation of the value obtained by solving the fdP equation [28]. Therefore, efficient and reasonably accurate procedures for the determination of effective Born radii (or atomic solvation energies) can be used. In this area the most recent developments of GB models have taken place [8]. The standard approach starts by assuming the so-called *Coulomb field approximation*, where the electric displacement $\overrightarrow{D}_i$ for each atom $i$ is calculated by supposing that the dielectric boundary is spherical and that atom $i$ lies at the center of this sphere. (Note that this spherical symmetry is only assumed to calculate $\overrightarrow{D}_i$). A large variety of procedures for calculating effective Born radii within the Coulomb field approximation have been presented. These include numerical surface or volume integration [16, 21, 25, 29, 30], analytical integral expression [22], and pairwise summation approximations [31, 32, 33].

Recently, corrections to the Coulomb field approximation have been suggested and shown to greatly increase the accuracy of the effective Born radii [21, 30, 34]. In a different approach it was demonstrated that the quantity $\sqrt{R_i R_j}$ can be interpreted as a measure of enclosure of the $(i, j)$ atom pair and be calculated very efficiently to yield accurate screened interaction energies [23]. The development of FACTS (see next subsection) was inspired by this measure of enclosure.

### 2.3 Fast Analytical Continuum Treatment of Solvation (FACTS)

In FACTS, the self electrostatic solvation energy and SASA of individual atoms are calculated using intuitive geometric properties of the solute whose evaluation requires only solute interatomic vectors. For each solute atom the volume and spatial symmetry of its neighboring atoms or, equivalently, of the solvent displaced by the neighboring atoms, are approximated. A linear combination with cross-term of these two measures is used as independent variable of a sigmoidal function (see below). The parameters of the sigmoidal function, together with those of the linear combination with cross-term, are derived by fitting to atomic electrostatic solvation energy values calculated by numerical solution of the fdP equation. The GB formula (6) is used to obtain the electrostatic solvation free energy of the macromolecule. The FACTS model does not assume the Coulomb field approximation (see above) and does not require to define a dielectric discontinuity surface. (Such dielectric boundary is only required to calculate the fdP reference data to which the parameters of the FACTS model are fitted.)

The same two measures of solvent displacement are combined and used in another sigmoidal function to estimate the SASA of individual atoms. The parameters of the sigmoidal function are derived by fitting to SASA values calculated by an exact analytical method [35]. Finally, the nonpolar contribution to the solvation free energy is assumed to be proportional to the sum of the atomic SASA values [36, 37].

Both electrostatic solvation energy and SASA are determined using the same geometrical properties and analytical framework, which makes FACTS a comprehensive and efficient implicit solvation model.

### 2.3.1 Atomic (or Self) Electrostatic Solvation Energy

The essential idea in FACTS is that the electrostatic solvation free energy of atom $i$, $\Delta G_i^{el}$, is evaluated by considering a sphere of radius $R_i^{sphere}$ around atom $i$. The radius is large enough to neglect effects on $\Delta G_i^{el}$ due to conformational changes outside the sphere. (More precisely, let $\Delta G_i^{el,m}$ denote $\Delta G_i^{el}$ being calculated with the region outside of $R_i^{sphere}$ occupied exclusively by atoms of the macromolecule, assuming an infinitely large protein. Similarly, let $\Delta G_i^{el,s}$ denote $\Delta G_i^{el}$ being calculated with the region outside of $R_i^{sphere}$ occupied exclusively by solvent. Then the value of $R_i^{sphere}$ is chosen large enough so that $\Delta G_i^{el,s} - \Delta G_i^{el,m} \cong 0$ holds for any conformation within the sphere.) If only atom $i$ of the macromolecule were present within the sphere of radius $R_i^{sphere}$, solving the Poisson equation would result in $\Delta G_i^{el} \cong -\frac{\tau q_i^2}{2r_i^{vdW}}$. As more and more atoms are gradually added (see Figure 1) $\Delta G_i^{el}$ becomes less favorable depending in a complex way on where the additional atoms are placed. When all the solvent has finally been flushed out from within the sphere, solving the Poisson equation would result in $\Delta G_i^{el} \cong 0$.

To quantify the atomic solvation energy, it is useful to investigate the change in solvation energy upon sequential addition of solute atoms to the interior of the sphere. Two desolvation pathways are shown in Figure 1. In the leftmost column the atom at the center is completely solvated. In the rightmost column it is completely desolvated. In proceeding from left to right on the top or bottom row in Figure 1, more and more atoms surrounding the atom at the center are added. Thus, the central atom becomes more and more desolvated and its solvation energy $\Delta G_i^{el}$ becomes less and less favorable. The difference between the two pathways is that on the top pathway, atoms are added such as to disrupt the spatial symmetry within the sphere, whereas on the bottom pathway atoms are added such as to preserve the spatial symmetry. Crossing from the asymmetric to the symmetric pathway in the two intermediate steps in Figure 1, i.e., going from $b$ to $f$ and $c$ to $g$, respectively, the number of atoms surrounding the central atom remains constant but they are rearranged such that solvent closer to the central atom is displaced. Thus the solvation energy of the central atom becomes less favorable. The following two observations are the core of the FACTS model. The increase in solvation energy induced by adding solute atoms (from left to right in Figure 1) can be accounted for by the change of a suitably defined measure of volume. It quantifies the volume occupied by solute atoms within the sphere of radius $R_i^{sphere}$. The increase in solvation energy originating from a rearrangement of solute atoms (from top to bottom in Figure 1) can be approximated by the change of a measure of symmetry, which quantifies the symmetry of the spatial distribution of the atoms surrounding atom $i$.

From the previous description it is clear that a measure of volume or symmetry alone is not appropriate to calculate the solvation energy of atom $i$. A fully buried atom (Figure 1*d*) and a fully exposed atom (Figure 1*a*) are only marginally discriminated by the spatial symmetry within the sphere. (However, the latter situation never arises in proteins since each atom always has neighbors.) Hence, the number of neighbors is the key difference. Analogously, the volume occupied by solute atoms within the sphere is constant in, for instance, snapshots $b$ and $f$ in Figure 1. Nevertheless, the solvation energy becomes less favorable by crossing from $b$ to $f$. In this case the key difference is the symmetry. Either of the two measures provides a partial description, but a synergistic combination

of the two measures yields a powerful means to calculate the atomic solvation energy.

To cast the above ideas into a mathematical form, the abbreviations $\overrightarrow{x}_{ij} = \overrightarrow{x}_i - \overrightarrow{x}_j$, $r_{ij} = |\overrightarrow{x}_{ij}|$, and $\hat{x}_{ij} = \frac{\overrightarrow{x}_{ij}}{r_{ij}}$ are introduced. The measure of *volume* occupied by the solute around atom $i$ is defined by

$$A_i = \sum_{j=1, j\neq i}^{N} V_j \Theta_{ij} \tag{7}$$

and the measure of *symmetry* by

$$B_i = \left| \frac{\sum_{j=1, j\neq i}^{N} \frac{V_j}{r_{ij}} \Theta_{ij} \hat{x}_{ij}}{\sum_{j=1, j\neq i}^{N} \frac{V_j}{r_{ij}} \Theta_{ij}} \right| \tag{8}$$

where

$$\Theta_{ij} := \begin{cases} \left(1 - \left(\frac{r_{ij}}{R_i^{sphere}}\right)^2\right)^2 & r_{ij} \leq R_i^{sphere} \\ 0 & r_{ij} > R_i^{sphere} \end{cases} \tag{9}$$

The measure of volume $A_i$ is simply the sum of the van der Waals volumes $V_j$ of the atoms surrounding atom $i$ within the sphere, weighted by $\Theta_{ij}$. Typically $A_i$ ranges between 100 Å$^3$ and 2000 Å$^3$ in a sphere of radius $R_i^{sphere} \cong 10$ Å. The measure of symmetry $B_i$ is a weighted Euclidean norm of the sum of the unit vectors pointing from the central atom $i$ to the neighboring atoms. Thereby each unit vector is weighted by $\Theta_{ij}$, and additionally by the volume of the neighboring atom $V_j$ it points to, divided by its distance $r_{ij}$ from atom $i$. There is no other reason for the additional weighting factor $V_j/r_{ij}$ except for the fact that it was found to improve the correlation between the values of $B_i$ and atomic solvation energies calculated by fdP. The value of $B_i$ is normalized to range between 0 and 1. For a fully symmetric distribution $B_i = 0$, whereas for a totally asymmetric distribution (e.g., only one neighboring atom) $B_i = 1$. The purpose of the function $\Theta_{ij}$ is twofold: weighting and smoothing. $\Theta_{ij}$ is equal to 1 for $r_{ij} = 0$ and drops continuously until $\Theta_{ij} = 0$ at $r_{ij} = R_i^{sphere}$. Thus, on the one hand $\Theta_{ij}$ accounts for the fact that the further an atom is placed from atom $i$, the less it influences its solvation energy. On the other hand $\Theta_{ij}$ ensures the existence of continuous (first and second) derivatives. Note that due to the function $\Theta_{ij}$ the measure of volume includes a small contribution originating from the symmetry. As an example, configurations $c$ and $g$ in Figure 1 yield different values for the measure of volume because of $\Theta_{ij}$.

Having defined the measure of volume and symmetry, the next step is to obtain a functional relationship between atomic solvation energies and the quantities $A_i$ and $B_i$. The aim is to find a function prototype with some parameters that can be optimized to reproduce accurately the fdP reference values. At this point it is important to note that once a function prototype is found, its parameters have to be optimized separately for each van der Waals radius of the solute atoms. To explain the importance of the van der Waals radius one can consider two fully solvated atoms with differing van der Waals radii. The two atoms have the same values for the measure of volume (zero) and symmetry (zero), but their solvation energies are different and depend on the van der Waals radii according to

the Born formula.

To obtain the desired relationship it is helpful to plot fdP derived atomic solvation energies $\Delta G_i^{el,fdP}$ for unit charges against $A_i$ and $B_i$ in a three-dimensional graph (Figure 2), classified in sets according to the van der Waals radii of the corresponding atoms. For each set a sigmoidal distribution of data is observed. Therefore, the measures of volume and symmetry are combined linearly and by a mixed term into a single measure of solvent displacement

$$C_i = A_i + b_1 B_i + b_2 A_i B_i \tag{10}$$

and a sigmoidal shaped function of $C_i$ is used to calculate the electrostatic solvation energy $\Delta G_i^{el,FACTS}$ of atom $i$ for a unit charge:

$$\Delta G_i^{el,FACTS} = a_0 + \frac{a_1}{1 + e^{-a_2(C_i - a_3)}} \tag{11}$$

The parameters $a_0$ and $a_1$ are determined using the limiting cases of a fully buried and fully exposed atom. In the case of a fully buried atom (i.e., $C_i \to +\infty$) the value of $\Delta G_i^{el}$ should vanish which implies that $a_0 = -a_1$ and $a_2 > 0$. For a fully exposed atom (i.e., $C_i \to -\infty$) the Born formula applies so that $a_0 = -\frac{\tau q_i^2}{2r_i^{vdW}}$. Hence, for each van der Waals radius the five parameters $b_1$, $b_2$, $a_2$, $a_3$, and $R^{sphere}$ have to be determined by an optimization procedure. The sigmoidal function (equation 11) gives an accurate fit to $\Delta G_i^{el,fdP}$ (Figure 2). Intuitively, $C_i$ measures the solvent displacement around atom $i$, and the solvation energy of atom $i$ is a sigmoidal function of this measure. Using the definition of effective Born radius given in equation (3),

$$R_i^{FACTS} = -\frac{\tau q_i^2}{2\Delta G_i^{el,FACTS}} \tag{12}$$

and the GB formula for the interaction term (i.e., the second sum on the r.h.s. of equation (5)), the total electrostatic solvation energy in the FACTS model is written as

$$\Delta G^{el,FACTS} = \sum_{i=1}^{N} \Delta G_i^{el,FACTS} - \tau \sum_{1 \le i < j \le N} \frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i^{FACTS} R_j^{FACTS} \exp(-r_{ij}^2 / \kappa R_i^{FACTS} R_j^{FACTS})}} \tag{13}$$

$$= -\frac{\tau}{2} \sum_{i,j=1}^{N} \frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i^{FACTS} R_j^{FACTS} \exp(-r_{ij}^2 / \kappa R_i^{FACTS} R_j^{FACTS})}} \tag{14}$$

where $r_{ii} = 0$ and $N$ is the number of atoms in the macromolecule. Note that the second sum in equation (13) implies an infinite cutoff while a truncation scheme (shifting [38]) is used in the MD simulations reported below.

### 2.3.2 Atomic Solvent Accessible Surface Area

Estimating amount and symmetry of the solvent that is displaced around a given atom provides information on how much the atom is accessible to solvent. Therefore, the geometric concepts described above for approximating the atomic electrostatic solvation energy can also be used to calculate the SASA. Several efficient methods that accomplish this task have been suggested in the past. They mainly use interatomic distances only and do not take into account symmetry. It has been suggested that angles between atom triplets could be used [39], but such an approach is too time consuming. The FACTS approach offers a straightforward way to approximate the SASA of atom $i$, $S_i$, by taking into account the relative positions of the surrounding atoms. Analogously to equation (10) one can define

$$D_i = A_i + d_1 B_i + d_2 A_i B_i \tag{15}$$

and

$$S_i^{FACTS} = c_0 + \frac{c_1}{1 + e^{-c_2(D_i - c_3)}} \tag{16}$$

for the SASA of atom $i$. The parameters $c_0$ and $c_1$ are determined using the limiting cases of a fully buried and fully exposed atom. In the case of a fully buried atom (i.e., $D_i \to +\infty$) the value of $S_i$ should vanish which implies that $c_0 = -c_1$ and $c_2 > 0$. For a fully exposed atom (i.e., $D_i \to -\infty$) the analytical formula applies so that $c_0 = 4\pi(r_i^{vdW} + 1.4)^2$ using a probe sphere of 1.4 Å radius. The parameters $d_1$, $d_2$, $c_2$ and $c_3$, are derived by fitting to exact values of the SASA [35].

### 2.3.3 Total Solvation Free Energy in the FACTS Model

The solvation free energy of a macromolecule is written as the sum of a polar and a non-polar term

$$\Delta G^{FACTS} = \Delta G^{el,FACTS} + \gamma \sum_{i=1}^{N} S_i^{FACTS} \tag{17}$$

where $\Delta G^{el,FACTS}$ is detailed in equation (14), and $\gamma$ denotes the empirical surface tension parameter. Values of $\gamma = 0.015$ and $\gamma = 0.025$ kcal mol$^{-1}$ Å$^{-2}$ were used for the MD simulations presented in the Results section.

## 2.4 Parameterization of FACTS

### 2.4.1 Peptides and Proteins

A composite set of 5 structured peptides (1cb3, a $\beta$-sheet from 1pgb, an $\alpha$-helix from 1pgb, 1ly2, and Beta3s [40]), 18 single-chain proteins (1a2p, 1bpi, 1crn, 1dvd, 1f8a, 1fmk, 1hdn, 1h0l, 1inc, 1lz1, 1pgb, 1pht, 1shg, 1ubq, 2ci2, 2ptl, 3app, and 3pte), and 6 multi chain proteins (1kvd, 1ycq, 1ycr, 2ins (chains A and B), 2ins (all chains), and

5hvp) of very different sizes, shapes, and secondary structure content was used. The number of residues ranges from 11 in 1cb3 to 347 in 3pte. The set includes almost spherical geometries with no cavities as well as structures with internal cavities. For instance, 5hvp is the HIV-1 aspartic proteinase in a complex with a peptidic ligand that was removed from the active site to obtain an internal cavity. To further diversify the set of structures with many different kinds of irregular shapes (cavities, open loops, etc.) the single chain proteins were subjected to high temperature unfolding simulations at 450 K for 20 ns with an implicit solvent model [19]. From each trajectory a molten globule-like structure and a significantly extended conformation were selected and added to the initial set of structures. (For 1bpi only a molten globule-like structure was chosen as it is strongly stabilized by three disulfide bridges and did not unfold sufficiently in the simulation. Similarly, for the very large complexes 1inc, 3app, and 3pte only a molten globule-like structure was added as a significantly extended conformation is too memory demanding for the fdP calculations.) The average increase in the radius of gyration is 26.1% and 92.2% for the molten globule like and significantly unfolded structures, respectively. Their average $C_\alpha$-RMSD (root mean square deviation) is 13.1 Å and 17.7 Å, respectively. Furthermore, almost completely extended conformations of the structured peptides were included. The final training set consists of 81 (PARAM19, see below) and 72 (PARAM22, see below) conformations from 29 peptides and proteins.

### 2.4.2 Small Molecules

Recently, the potentials of mean force between pairs of charged side chains have been calculated in explicit water [41]. From this study a total of 12 arrangements originating from 7 molecular systems were selected: Glu-Glu head to head and orthogonal, His-Glu orthogonal, Lys-Glu head to head and orthogonal, Lys-Lys head to head, Arg-Glu head to head, Arg-Lys head to head and orthogonal, Arg-Arg head to head, orthogonal and stacked. The distance was varied from 2.4 Å to 10 Å resulting in 77 conformations for each pair. Detailed descriptions of the structures and definitions of the distances are given in [41]. Furthermore, 77 conformations of the N-methyl-acetamide dimer in a planar arrangement were also considered. Again, the distance between the hydrogen bond donor and acceptor atoms was varied from 2.4 Å to 10 Å. The union set of all peptides, proteins, and small molecules consists of 1082 structures (81 protein conformations, 77x12 arrangements of pairs of charged side chains, and 77 N-methyl-acetamide dimer arrangements) derived from 37 molecular systems (29 proteins, 7 pairs of charged side chains, and the N-methyl-acetamide dimer). This constitutes a sound basis for a thorough fitting and assessment of the FACTS model.

### 2.4.3 Force Field Parameter Set

All calculations were performed using the CHARMM program (version c29b1) with the CHARMM polar hydrogen parameter set (PARAM19 [38]) and the CHARMM all-hydrogen parameter set (PARAM22 [42]). For the PARAM19 set the van de Waals radii of all hydrogen atoms are set to 1 Å in the fdP, FACTS, and GBMV calculations. For

some computations (e.g., atomic solvation energies) all atoms are assigned unit charges to allow for a comparison that is unbiased by the charge parameter set.

### 2.4.4 Finite difference Poisson (fdP)

The benchmark commonly used to assess the accuracy of continuum electrostatics models are the energy values calculated by fdP. Atomic solvation energies $\Delta G_i^{el,fdP}$ and pair interaction energies were calculated by numerical solution of the fdP equation with the PBEQ module [43] in CHARMM. All atoms were assigned unit charges for the fdP calculations. A grid spacing of 0.2 Å was used for all fdP calculations with proteins. For the pairs of charged side chains and the N-methyl-acetamide dimer a grid spacing of 0.1 Å was used. The van der Waals radii of all hydrogen atoms were set to 1 Å for PARAM19. No adjustments were applied to the van der Waals radii of PARAM22. The dielectric discontinuity boundary was defined by the molecular surface. The atomic solvation energy $\Delta G_i^{el,fdP}$ of atom $i$ is the solvation energy of the macromolecule when deleting the charges of all atoms except the one of atom $i$. Solvation energies were calculated by subtracting the self energy in vacuo ($\varepsilon_m = 1$, $\varepsilon_s = 1$) from the self energy in solution ($\varepsilon_m = 1$, $\varepsilon_s = 78.5$). The interaction energy of an $(i, j)$ atom pair was obtained by calculating the electrostatic energy of a unit charge at the position of atom $j$ in the electric field generated by a single unit charge at the position of atom $i$ in the presence of solvent ($\varepsilon_m = 1$, $\varepsilon_s = 78.5$).

### 2.4.5 Parameter Optimization

For each van der Waals radius two sets of parameters have to be optimized separately: the five parameters $b_1$, $b_2$, $a_2$, $a_3$, and $R^{sphere}$ for the atomic solvation energies, and the four parameters $d_1$, $d_2$, $c_2$, and $c_3$ for the atomic SASA. Note that an upper bound of 10 Å was imposed for the optimization of $R^{sphere}$. Furthermore, $R^{sphere}$ was optimized only for electrostatic solvation energies. For atomic SASA values the $R^{sphere}$ parameters determined for the electrostatic solvation are used to increase efficiency in MD simulations as the same atom-pair list can be used for the electrostatic solvation energy and SASA. Optimal parameters were obtained by minimizing the deviations of $\Delta G_i^{el,FACTS}$ from $\Delta G_i^{el,fdP}$ and of $S_i^{FACTS}$ from $S_i^{exact}$. A particle swarm algorithm [44] was used for parameter optimization. The fdP data from the 81 conformations (of the 29 peptides and proteins listed above) are included in the training set. The data for the charged side chain pairs and the N-methyl-acetamide dimer are only used for tests. All parameters are listed in the Suppl. Mat.

It is interesting and useful to assess the dependency of the FACTS parameters on the training set. The dependency is marginal because fitting on a single medium sized and globular protein (e.g., the native state of barnase (PDB code 1a2p)) yields a parameter set that performs almost as good as using all 81 protein conformations (Table 1 and Suppl. Mat). The only protocol that fails to produce reliable parameters is to fit only on small or very extended conformations. In both these cases the radius $R^{sphere}$ is estimated too small, resulting in a significant loss of accuracy for large and compact conformations. On the other hand, a larger radius does not compromise the

accuracy for small or very open structures (but has a negative effect on the efficiency). In retrospect these findings show that the data set used in this study to obtain the FACTS parameters is redundant. Yet, these findings are useful for additional parameterizations of the FACTS model (e.g., for CHARMM and $\varepsilon_m = 4$ or for another force field), which can be done with much less fdP data and therefore much faster.

### 2.5 MD Simulations

All MD simulations were performed with CHARMM [38] starting from the native structure downloaded from the PDB database [45]. Constant temperature MD simulations were carried out using weak coupling to a Berendsen's bath with a coupling constant of 5 ps. The CHARMM default truncation schemes of long-range electrostatics were used, i.e., a shift to zero energy at 7.5 Å and 12 Å for PARAM19 and PARAM22, respectively. The same cutoff values were employed for the van der Waals energy with a shifting and polynomial switching function for PARAM19 and PARAM22, respectively. The SHAKE algorithm was used to fix the length of the covalent bonds involving hydrogen atoms, which allows an integration time step of 2 fs. The non-bonding interactions were updated using a heuristic update algorithm and coordinate frames were saved every 10 ps for analysis.

## 3 Results and Discussion

This section focuses on the results obtained using an interior (i.e., solute) dielectric $\varepsilon_m = 1$. Note that using $\varepsilon_m = 1$ (instead of $\varepsilon_m = 2$ or $\varepsilon_m = 4$) is the most stringent test of the accuracy of a continuum dielectric model. For single point energy calculations (e.g., for structure prediction or ranking in ligand binding) $\varepsilon_m = 2$ or $\varepsilon_m = 4$ would be more appropriate since values of $\varepsilon_m > 1$ account for thermal fluctuations of protein dipoles. As the FACTS model is primarily aimed to be used in MD simulations, the validation with $\varepsilon_m = 1$ is discussed in detail in the present study. However, parameterizations of the FACTS model for $\varepsilon_m = 2$ have also been performed and results are presented in the Suppl. Mat.

### 3.1 Atomic (or Self) Electrostatic Solvation Energy

It is interesting to assess the gain in accuracy by combining the measures of volume and symmetry instead of using only either of them. For this purpose the optimizing procedure for atomic electrostatic solvation energies was performed three times: by using both measures (as in equation (10), i.e., $C_i = A_i + b_1 B_i + b_2 A_i B_i$), by using only the measure of volume ($\tilde{C}_i = A_i + \tilde{b}_1 A_i^2 + \tilde{b}_2 A_i^3$), and by using only the measure of symmetry ($\hat{C}_i = B_i + \hat{b}_1 B_i^2 + \hat{b}_2 B_i^3$). Note that the number of parameters is the same in all three situations. Plots of atomic solvation energy values calculated with the three different $C_i$'s versus fdP values are shown in Figure 3. Interestingly, the measure of volume yields more accurate solvation energies than the measure of symmetry for buried atoms (solvation energy close to zero), whereas the measure of symmetry is better for solvent exposed atoms (favorable solvation energy).

This observation provides evidence for the synergistic effect of combining the two measures.

Figure 4 shows atomic electrostatic solvation energy values calculated by FACTS (equation (11)), GBMV2 [21], and GBMVgrid [30] versus the benchmark fdP values. The numerical approach GBMVgrid is the most accurate method, followed by GBMV2 and FACTS. However, the maximal absolute error is largest for GBMV2 because of some significant outliers. Similar behavior is observed for both PARAM19 and PARAM22.

### 3.2 Atomic Solvent Accessible Surface Area (SASA)

The correlation between SASA values of atoms in protein structures calculated by FACTS and exact values is 0.96 and 0.97 for PARAM19 and PARAM22, respectively (Figure 5). The accuracy of the GBMV surface algorithm [21] is slightly higher than FACTS, and more so for PARAM19. The largest deviations in FACTS PARAM19 are observed for atoms with little solvent accessibility and originate from the relatively large sphere radii of the carbon atoms, which are close to 10 Å (Suppl. Mat.). It has to be remembered that the sphere radii were not optimized *ad hoc* for the atomic SASA evaluation but set equal to those of the electrostatic atomic solvation energy for computational efficiency. Large discrepancies are observed mainly for the small molecular systems, i.e., pairs of charged side chains and the N-methyl-acetamide dimer, which is also a consequence of the large sphere radii. Interestingly, both GBMV and FACTS yield more accurate atomic SASA values than the approach by Hasel et al. [39] (see Suppl. Mat.).

### 3.3 Pairwise Electrostatic Energies and Their Sums

Screened interaction energies, i.e., pairwise energies in solution [46], are calculated for FACTS, GBMV2, and GBMVgrid by the formula

$$G_{ij} = \frac{q_i q_j}{\epsilon_m r_{ij}} - \frac{\tau q_i q_j}{\sqrt{r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/\kappa R_i R_j)}} \tag{18}$$

where the Born radii $R_i$ are evaluated using the respective models. The agreement with fdP values is excellent for all three methods (Suppl. Mat.). Yet, in MD simulations one is not primarily interested in individual pair interaction energies $G_{ij}$. The relevant quantity for MD is the sum over all pairwise energies involving a given solute atom $i$, i.e., $G_i = \sum_{j \neq i} G_{ij}$, because this quantity determines the contribution to the force on atom $i$ that is due to the electrostatic interaction. Accurate reproduction of $G_{ij}$ in a given model with respect to the fdP values does not necessarily imply accurate reproduction of $G_i$ since individual errors may not compensate among each other. A good agreement with fdP values of $G_i$ is obtained using FACTS (Suppl. Mat.).

### 3.4 Electrostatic Solvation Energy of Protein Conformations

To assess the accuracy of FACTS in calculating macromolecular solvation energy a large variety of protein conformations were generated by 50-ns MD simulations of unfolding at 450 K using an implicit solvent model [19]. Coordinates were saved every 10 ps and all snapshots were sorted according to increasing radius of gyration ($R_g$). A total of 100 conformations were chosen from each trajectory as follows: every 20th conformation from the 500 snapshots with the lowest $R_g$ (25 conformations), every 20th conformation from the 500 snapshots with the largest $R_g$ (25 conformations), and every 80th conformation from the remaining 4000 snapshots (50 conformations). The 100 conformations of each protein cover a wide range of RMSD and $R_g$. For the $29 \times 100$ conformations the values of the electrostatic solvation energy $\Delta G^{FACTS}$ (equation(13)), $\Delta G^{GBMV2}$, and $\Delta G^{GBMVgrid}$ were calculated and compared to $\Delta G^{fdP}$.

The results for barnase (1a2p) show that the agreement between approximated and exact (i.e., fdP) values is very good for the three models (Figure 6). As indicated by the percentage error for all proteins, the accuracy of FACTS improves significantly by using $\kappa = 8$ instead of $\kappa = 4$, while only a marginal improvement is observed with $\kappa = 12$ (Table 2). Notably, with $\kappa = 12$ the percentage error of $\Delta G^{FACTS}$ averaged over all 2900 conformations is only 1.36% and 1.56% with PARAM19 and PARAM22, respectively. Only with PARAM22 is GBMV2 (with its default value of $\kappa = 8$ [21]) more accurate than FACTS, which is probably a consequence of the fact that GBMV2 was optimized mainly for PARAM22. The cumulative histogram (Figure 7) shows that 95% of the 2900 conformations have an error in the FACTS solvation energy smaller than 3.66% and 3.72% with PARAM19 and PARAM22, respectively.

For most applications of force-field based methods, the crucial quantity is the *difference* in electrostatic solvation energy between two structures of the same molecular system, i.e., $\Delta\Delta G$. These differences are calculated for all pairs of structures for each trajectory for FACTS, GBMV2 and GBMVgrid and compared to the fdP values. The results are shown in Figure 8 and Table 3. FACTS performs almost as well as GBMV2. The GBMVgrid approach is the most accurate of the three models but cannot be used for MD simulations because it is a numerical method.

### 3.5 Molecular Dynamics Simulations

The FACTS implementation into CHARMM version c29b1 has passed "TEST FIRST", which is a stringent check of first derivatives by a comparison with the numerical (i.e., finite difference) evaluation of the gradient. Most importantly, the total energy does not drift in NVE simulations even with a time step of 2 fs (Figure 9), whereas GBMV requires a time step of 1 fs to reduce the energy drift [47].

The native state of structured peptides and proteins is stable over 100-ns MD runs at 300 K (Table 4). Interestingly, the MD results are similar for two different values of the surface tension-like parameter ($\gamma = 0.015$ and $\gamma = 0.025$ kcal mol$^{-1}$ Å$^{-2}$), which indicates robustness with respect to the relative weighting (i.e., balancing) of polar and nonpolar solvation. Only 1cb3 and 1abz show a $C_\alpha$-RMSD larger than 3.5 Å after 100 ns with both values

of the parameter $\gamma$. These findings are consistent with experimental data. The PDB entry 1cb3 is an ensemble of NMR conformers of the segment 101-111 of $\alpha$-lactalbumin, which is flexible when isolated from the context of the protein. In fact, the five C-terminal residues of this segment were shown by NMR to be essentially unstructured in water at 283 K [48]. The de novo designed 38-residue $\alpha$-helical hairpin peptide $\alpha t\alpha$ (PDB 1abz) was estimated to be only 60% helical at 298 K by circular dichroism [49].

A common artifact of MD simulations in vacuo is the very small atomic fluctuations. The RMS fluctuations of the $C_\alpha$ atoms of chymotrypsin inhibitor 2 (PDB 2ci2) along a FACTS PARAM22 300 K MD simulation are in agreement with the corresponding values derived from crystallographic B-factors (Figure 10). In particular, the N-terminal segment and the loop (residues 38-44) are the most flexible regions according to both MD simulations and X-ray data [50]. As expected, slightly larger fluctuations are observed with the smaller of the two values of the surface tension-like parameter $\gamma$ used for the non-polar term in the MD simulations.

The reversible folding to the NMR conformer has been observed in preliminary FACTS PARAM19 330 K simulations of Beta3s, a designed 20-residue three-stranded antiparallel $\beta$-sheet [40]. Moreover, the thermodynamic stability (i.e., free energy difference between folded and denatured state) of Beta3s is lower using FACTS than a SASA-based solvation model [19, 51], which is consistent with experimental data [52]. An in-depth analysis of reversible folding of structured peptides and small proteins will be presented elsewhere.

### 3.6 FACTS Computational Requirements

Using the same non-bonding cutoff, MD simulations with FACTS are nearly four times slower than in vacuo but about ten times faster than with GBMV2. Notably, on a single Opteron 1.8 GHz processor, a 100-ns MD run of the 46-residue crambin (1crn) requires 4 and 22 CPU-days with FACTS PARAM19 (396 atoms and 7.5 Å cutoff) and FACTS PARAM22 (642 atoms and 12 Å cutoff), respectively (Table 5). Moreover, the CPU-time scales linearly with protein size (Figure 11). The extra memory requirements for FACTS with respect to a vacuum calculation are marginal and they originate solely from the atom-pair list, which is used for both electrostatic and SASA calculations. As an example, with the current implementation of FACTS into the c29b1 version of CHARMM only 12 MBytes of RAM are needed for FACTS PARAM19 MD simulations of the 389-residue protein $\beta$-secretase (PDB 1sgz).

## 4 Conclusions

A fully analytical treatment of solvation in the continuum model has been presented. The method, called FACTS, is very efficient because it is based on simple measures of solvent displacement and thus requires only distances between solute atoms which are close in three-dimensional space. These interatomic distances have anyway to be calculated for the non-bonding terms of a force field. FACTS does not use a dielectric boundary nor does it assume the

Coulomb field approximation. The agreement between FACTS and numerical finite difference Poisson calculations is good and comparable to the one of the most accurate GB methods that introduce empirical corrections to the Coulomb field approximation. In MD simulations of proteins FACTS is about ten times faster than the most accurate GB implementations. The native state of structured peptides and proteins is stable during 300 K MD runs of 100 ns using FACTS in combination with the CHARMM force field. Moreover, marginally stable peptides and unstructured loops in proteins are flexible under the same conditions. The accuracy and efficiency of FACTS suggest that it could also be used for protein structure prediction and docking.

## 5 Acknowledgments

We thank Riccardo Pellarin, Andrea Prunotto, Manuel Walker, Nicolas Majeux, Giovanni Settanni and Francois Marchand for very interesting and useful discussions. The calculations were performed on a Beowulf Linux cluster at the Informatikdienste of the University of Zurich and we thank C. Bolliger, T. Steenbock and Dr. A. Godknecht for their help in setting up and maintaining the cluster. This work was supported by the Swiss National Science Foundation.

# References


[1] Tomasi, J. and Persico, M., *Chem. Rev.*, 1994, **94**, 2027–2094.

[2] Gilson, M. K., *Curr. Opin. Struct. Biol.*, 1995, **5**, 216–223.

[3] Roux, B. and Simonson, T., *Biophysical Chemistry*, 1999, **78**, 1–20.

[4] Cramer, C. J. and Trulhar, D. G., *Chem. Rev.*, 1999, **99**, 2161–2200.

[5] Bashford, D. and Case, D. A., *Annu. Rev. Phys. Chem.*, 2000, **51**, 129–152.

[6] Orozco, M. and Luque, F. J., *Chem. Rev.*, 2000, **100**, 4187–4225.

[7] Simonson, T., *Curr. Opin. Struct. Biol.*, 2001, **11**, 243–252.

[8] Feig, M. and Brooks III, C. L., *Curr. Opin. Struct. Biol.*, 2004, **14**, 217–224.

[9] Baker, N. A., *Curr. Opin. Struct. Biol.*, 2005, **15**, 137–143.

[10] Im, W.; Chen, J. and Brooks III, C. L., *Advan. Protein Chem.*, 2006, **72**, 171–195.

[11] Feig, M.; Chocholousova, J. and Tanizaki, S., *Theoretical Chemistry Accounts*, 2006, **116**, 194–205.

[12] Warwicker, J. and Watson, H. C., *J. Mol. Biol.*, 1982, **157**, 671–679.

[13] Gilson, M. K. and Honig, B. H., *Proteins: Structure, Function, and Bioinformatics*, 1988, **4**, 7–18.

[14] Bashford, D. and Karplus, M., *Biochemistry*, 1990, **29**, 10219–10225.

[15] Davis, M. E.; Madura, J. D.; Luty, B. A. and McCammon, J. A., *Comp. Phys. Comm.*, 1991, **62**, 187–197.

[16] Still, W. C.; Tempczyk, A.; Hawley, R. C. and Hendrickson, T., *J. Am. Chem. Soc.*, 1990, **112**, 6127–6129.

[17] Lazaridis, T. and Karplus, M., *Proteins: Structure, Function, and Bioinformatics*, 1999, **35**, 133–152.

[18] Hassan, S. A.; Guarnieri, F. and Mehler, E. L., *J. Phys. Chem. B*, 2000, **104**, 6478–.

[19] Ferrara, P.; Apostolakis, J. and Caflisch, A., *Proteins: Structure, Function, and Bioinformatics*, 2002, **46**, 24–33.

[20] Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A. and Brooks III, C. L., *J. Comput. Chem.*, 2003, **25**, 264–284.

[21] Lee, M. S.; Feig, M.; Salsbury, F. R. and Brooks III, C. L., *J. Comput. Chem.*, 2003, **24**, 1348–1356.

[22] Schaefer, M. and Karplus, M., *J. Phys. Chem.*, 1996, **100**, 1578–1599.

[23] Haberthür, U.; Majeux, N.; Werner, P. and Caflisch, A., *J. Comput. Chem.*, 2003, **24**, 1936–1949.

[24] Settanni, G.; Rao, F. and Caflisch, A., *Proc. Natl. Acad. Sci. USA.*, 2005, **102**, 628–633.

[25] Scarsi, M.; Apostolakis, J. and Caflisch, A., *J. Phys. Chem. A*, 1997, **101**, 8098–8106.

[26] Born, M., *Z. Phys.*, 1920, **1**, 45–48.

[27] Schaefer, M. and Froemmel, C., *J. Mol. Biol.*, 1990, **216**, 1045–1066.

[28] Onufriev, A.; Bashford, D. and Case, D. A., *J. Comput. Chem.*, 2002, **23**, 1297–1304.

[29] Ghosh, A.; Rapp, C. S. and Friesner, R. A., *J. Phys. Chem. B*, 1998, **102**, 10983–10990.

[30] Lee, M. S.; Salsbury, F. R. and Brooks III, C. L., *J. Chem. Phys.*, 2002, **116**, 10606–10614.

[31] Hawkins, G. D.; Cramer, C. J. and Trulhar, D. G., *J. Phys. Chem.*, 1996, **100**, 19824–19839.

[32] Qiu, D.; Shenkin, P. S.; Hollinger, F. P. and Still, W. C., *J. Phys. Chem. A*, 1997, **101**, 3005–3014.

[33] Dominy, B. N. and Brooks III, C. L., *J. Phys. Chem. B*, 1999, **103**, 3765–3773.

[34] Im, W.; Lee, M. S. and Brooks III, C. L., *J. Comput. Chem.*, 2003, **24**, 1691–1702.

[35] Lee, B. and Richards, F. M., *J. Mol. Biol.*, 1971, **55**, 379–400.

[36] Eisenberg, D. and McLachlan, A. D., *Nature*, 1986, **319**, 199–203.

[37] Ooi, T.; Oobatake, M.; Némethy, M. and Scheraga, H. A., *Proc. Natl. Acad. Sci. USA.*, 1987, **84**, 3086–3090.

[38] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S. and Karplus, M., *J. Comput. Chem.*, 1983, **4**, 187–217.

[39] Hasel, W.; Hendrickson, T. F. and Still, W. C., *Tetrahedron Comput. Methodol.*, 1988, **1**, 103–116.

[40] Ferrara, P. and Caflisch, A., *Proc. Natl. Acad. Sci. USA.*, 2000, **97**, 10780–10785.

[41] Masunov, A. and Lazaridis, T., *J. Am. Chem. Soc.*, 2003, **125**, 1722–1730.

[42] MacKerell Jr., A.D., e. a. and M., K., *J. Phys. Chem. B*, 1998, **102**, 3586–3616.

[43] Im, W.; Beglov, D. and Roux, B., *Computer Physics Communications*, 1998, **111**, 59–75.

[44] Kennedy, J. and Eberhart, R. C., *Swarm Intelligence*, Morgan Kaufmann Publishers, 2001.

[45] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N. and Bourne, P. E., *Nucl. Acids Res.*, 2000, **28**, 235–242.

[46] Scarsi, M. and Caflisch, A., *J. Comput. Chem.*, 1999, **14**, 1533–1536.

[47] Chocolousova, J. and Feig, M., *J. Comput. Chem.*, 2006, **27**, 719–729.

[48] Demarest, S. J.; Hua, Y. and Raleigh, D. P., *Biochemistry*, 1999, **38**, 7380–7387.

[49] Fezoui, Y.; Weaver, D. L. and Osterhout, J. J., *Proc. Natl. Acad. Sci. USA.*, 1994, **91**, 3675–3697.

[50] McPhalen, C. A. and James, M. N. G., *Biochemistry*, 1987, **26**, 261–269.

[51] Cavalli, A.; Haberthür, U.; Paci, E. and Caflisch, A., *Protein Science*, 2003, **12**, 1801–1803.

[52] De Alba, E.; Santoro, J.; Rico, M. and Jiménez, M. A., *Protein Science*, 1999, **8**, 854–865.

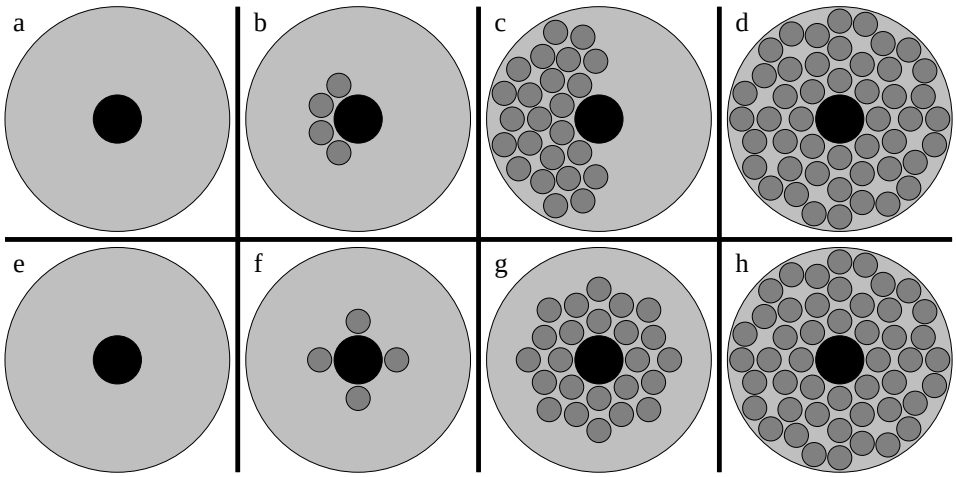


Figure 1: Schematic illustration of the essential concept of the FACTS evaluation of atomic solvation energy. The large circle in light gray represents the sphere of radius $R_i^{sphere}$ that is considered to quantify the atomic solvation energy in the FACTS approach (see text). The small circles in dark gray represent solute atoms that displace the solvent from around the central atom, which is in black. Both pathways ($a \rightarrow b \rightarrow c \rightarrow d$ and $e \rightarrow f \rightarrow g \rightarrow h$) proceed from a fully solvated to a fully desolvated atom. In the top pathway atoms are added such as to break spatial symmetry as much as possible. In the bottom pathway atoms are added such as to preserve spatial symmetry as much as possible. Crossing from the asymmetric (top) to the symmetric (bottom) pathway in the two intermediate steps, i.e., going from $b$ to $f$ or $c$ to $g$, the number of neighboring atoms remains constant but the solvation energy of the central atom changes significantly due to the increase in symmetry.

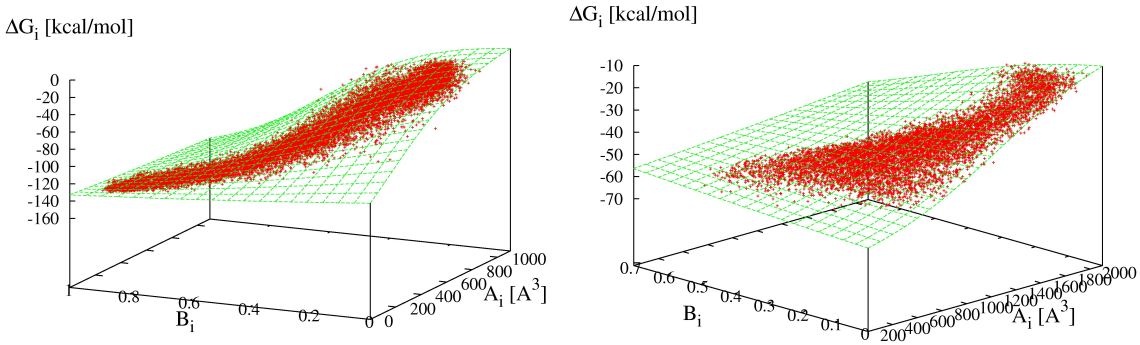


Figure 2: The green surface represents equation (11), i.e., FACTS atomic electrostatic solvation energy as a function of $A_i$ and $B_i$ for PARAM19 and a van der Waals radius of 1.0 Å (left) and 2.365 Å (right). The red data points are atomic solvation energy values calculated by fdP using unit charges and $\varepsilon_m = 1$. The dependence on the symmetry is more pronounced for the polar hydrogen atoms (left) than the carbon atoms (right) because the latter are usually more buried than the former. Note that a fully symmetric distribution yields $B_i = 0$.

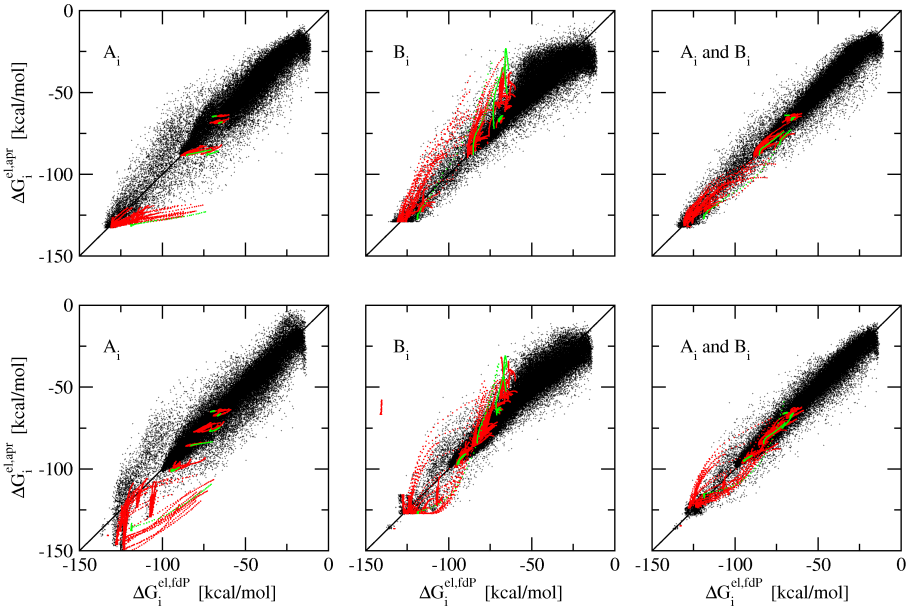


Figure 3: Synergistic effect of volume and symmetry terms in FACTS. In the left column only the measure of volume was used for the FACTS calculations, in the middle column only the measure of symmetry, and in the right column both measures were combined. Unit charges were used because they allow for a more stringent comparison that is not affected by the charge parameter set. The benchmark fdP calculations were performed with $\varepsilon_m = 1$. The data points of the protein conformations are in black, while those of the pairs of charged side chains and the N-methyl-acetamide dimer in red and green, respectively. (Top) Atomic electrostatic solvation energy values calculated by FACTS (equation (11)) versus the fdP values for 77'609 atoms from 1'082 molecular structures with the van der Waals radii of PARAM19. (Bottom) Atomic electrostatic solvation energy values calculated by FACTS (equation (11)) versus the fdP values for 90'747 atoms from 1'073 molecular structures with the van der Waals radii of PARAM22.

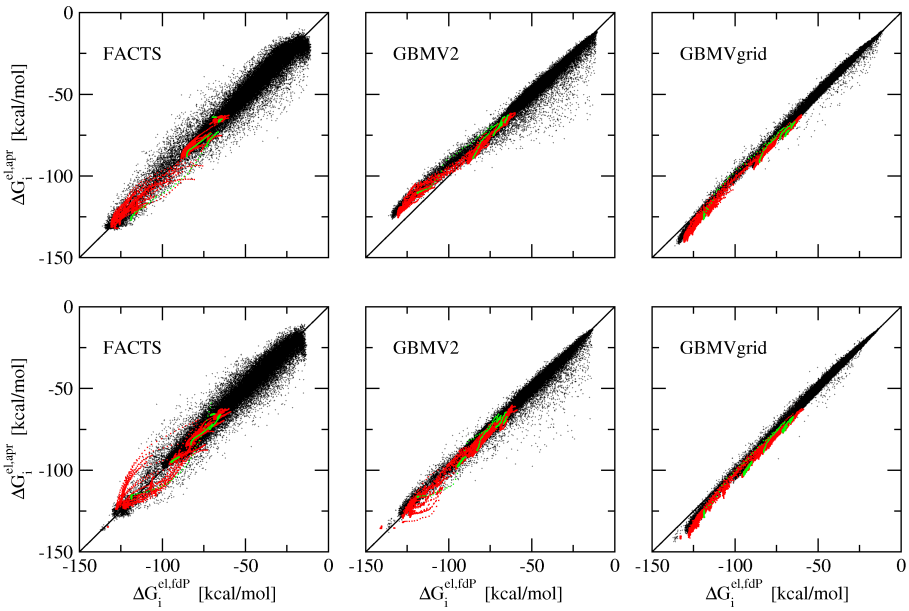


Figure 4: Comparison between FACTS and GBMV. The plots show values of the atomic electrostatic solvation energy evaluated with unit charges and $\varepsilon_m = 1$. The color coding for different molecular systems is the same as in Figure 3. (Top) PARAM19: Slope, correlation, and maximal absolute error for the 60'977 atoms in 81 protein structures are 0.963, 0.981, and 46.3 kcal/mol for FACTS; 0.910, 0.990, and 49.2 kcal/mol for GBMV2 [21]; 1.028, 0.998, and 23.0 kcal/mol for GBMVgrid [30]. (Bottom) PARAM22: Slope, correlation, and maximal absolute error for the 62'873 atoms in 72 protein structures are 0.964, 0.982, and 43.9 kcal/mol for FACTS; 0.967, 0.993, and 56.7 kcal/mol for GBMV2; 1.045, 0.998, and 19.2 kcal/mol for GBMVgrid.

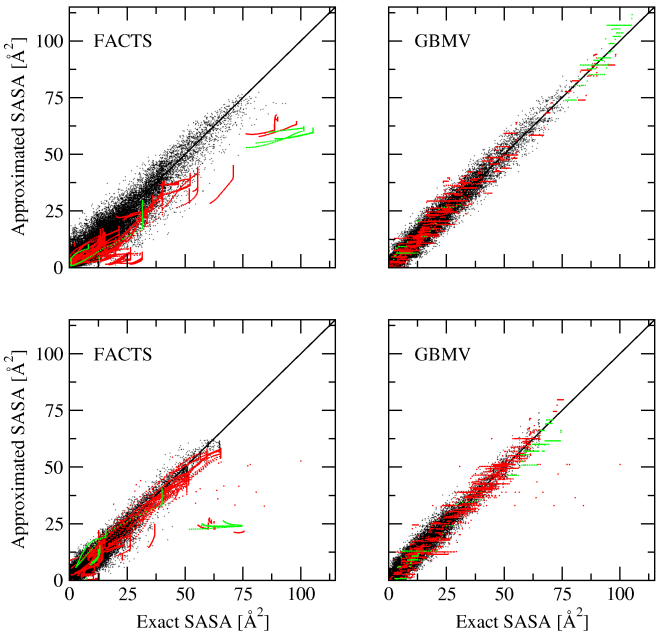


Figure 5: Comparison of atomic SASA evaluation by FACTS ($\varepsilon_m = 1$ parametrization) and GBMV [21]. The benchmark are the exact values of atomic SASA [35]. The color coding for different molecular systems is the same as in Figure 3. (Top) PARAM19: Slope, correlation, and maximal absolute error for the atoms in the protein structures are 0.915, 0.963, and 27.7 Å$^2$ for FACTS; 1.001, 0.986, and 14.1 Å$^2$ for GBMV. (Bottom) PARAM22: Slope, correlation, and maximal absolute error for the atoms in the protein structures are 0.939, 0.974, and 24.6 Å$^2$ for FACTS; 1.001, 0.983, and 14.3 Å$^2$ for GBMV.

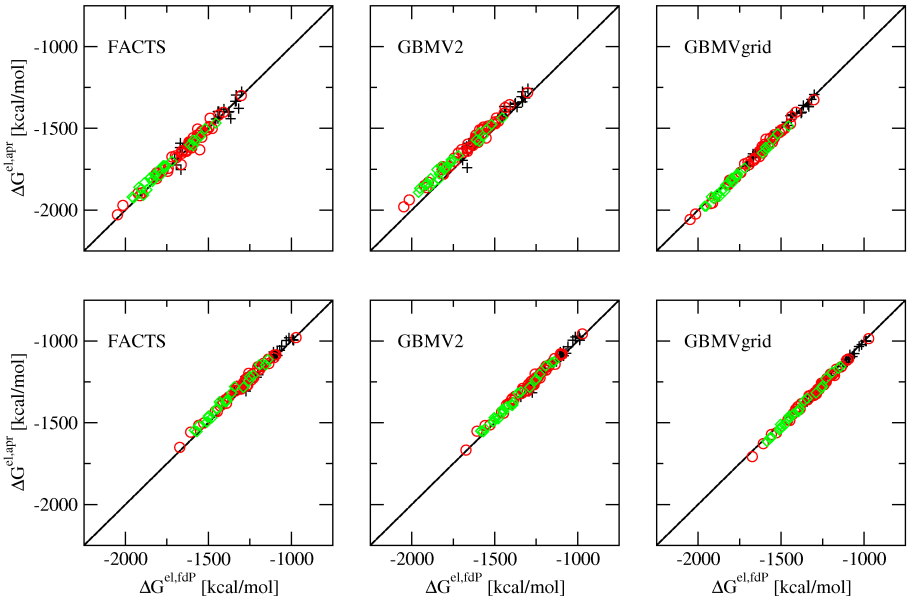


Figure 6: Comparison of protein electrostatic solvation energy values calculated by FACTS (equation (14)) and GBMV. Each plot shows data for 100 conformations of barnase with PARAM19 (top) and PARAM22 (bottom). The structures were chosen along a high temperature unfolding trajectory started from the 1a2p X-ray structure. Different symbols discriminate between different ranges of the radius of gyration. Plus and diamond symbols represent the 25 conformations with small and large radius of gyration, respectively, while circles the 50 intermediate ones. The benchmark are the fdP values with $\varepsilon_m = 1$.

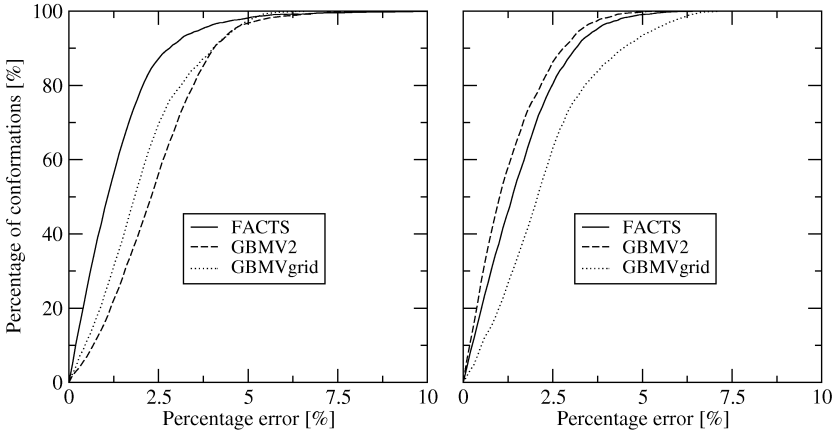


Figure 7: Cumulative histogram of percentage errors of electrostatic solvation energy values. The FACTS equation (14) was used for 2900 structures of 29 proteins for PARAM19 (left) and PARAM22 (right). The benchmark are the fdP values with $\varepsilon_m = 1$.

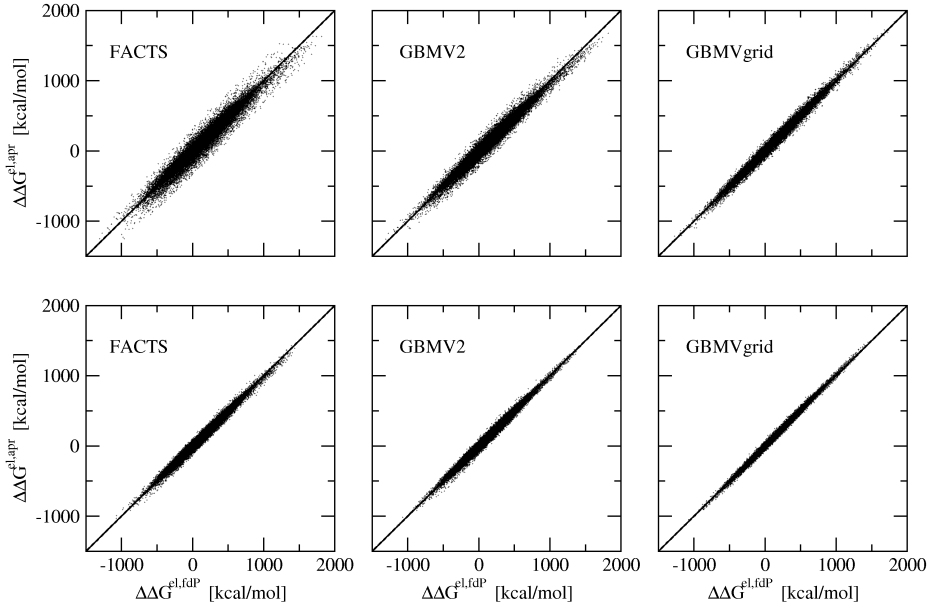


Figure 8: Comparison of relative electrostatic solvation energy values calculated by FACTS and GBMV ($\varepsilon_m = 1$). From each of 29 trajectories, 100 conformations were chosen as described in the text and the difference in solvation energy ($\Delta\Delta G$) for all possible pairs of structures was evaluated. In this way, an eventual systematic offset in solvation energy relative to the benchmark fdP values is eliminated. Such offset is irrelevant for MD simulations.

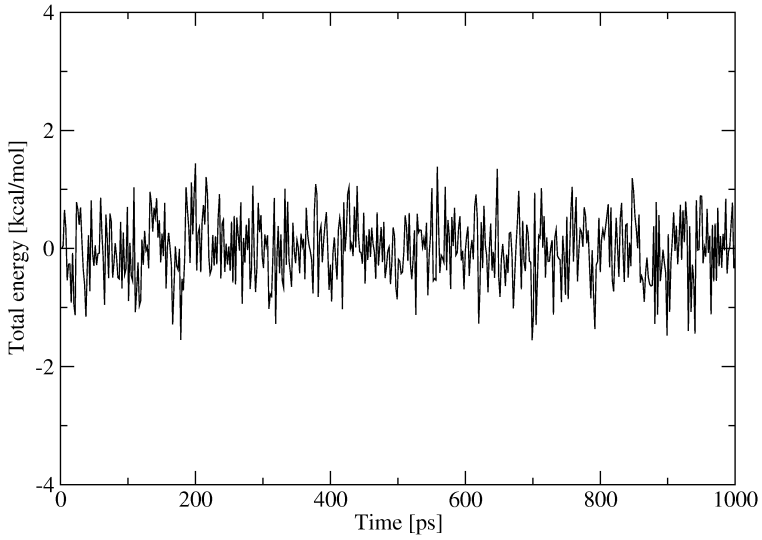


Figure 9: Time series of the total energy relative to the starting conformation of protein G (1pgb) during 1 ns MD simulation in the NVE ensemble using FACTS PARAM19 and a time step of 2 fs.

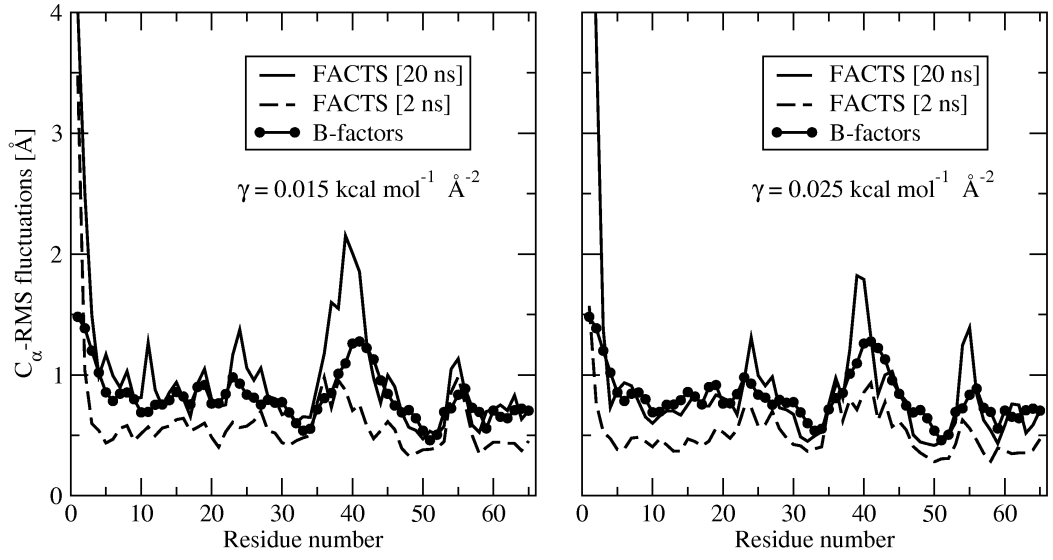


Figure 10: RMS fluctuations in Å of the $C_\alpha$ atoms of CI2. FACTS PARAM22 was used with $\varepsilon_m = 1$, and surface tension-like parameter $\gamma = 0.015$ kcal mol$^{-1}$ Å$^{-2}$ (left), and $\gamma = 0.025$ kcal mol$^{-1}$ Å$^{-2}$ (right). The fluctuations were extracted from a 300 K simulation started from the native structure (2ci2) and considering a trajectory segment of 2 ns (solid line) and 20 ns (dashed line). The bold line with circles represents the fluctuations derived from the crystallographic B-factors [50] using the formula RMS fluctuation = $(3B/(8\pi^2))^{0.5}$.

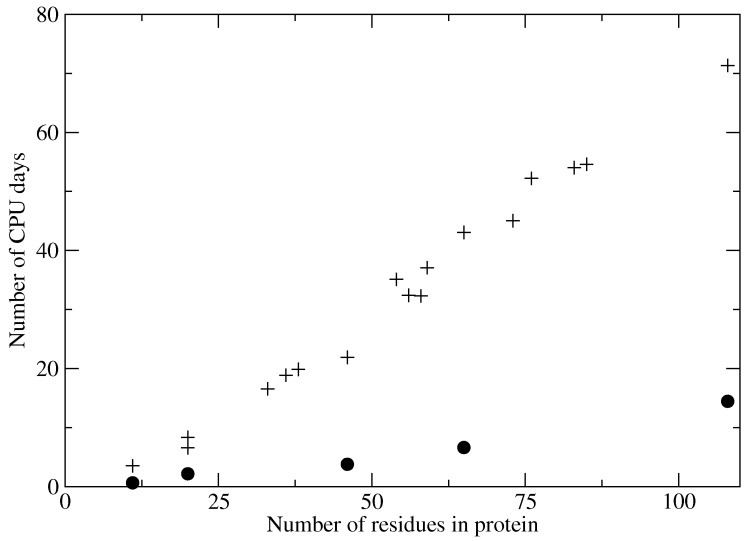


Figure 11: System-size scaling of CPU-time required for 100-ns MD simulations with FACTS. Circles and plus symbols correspond to simulations with PARAM19 and PARAM22, respectively.

| | training set: all<br>test set: all | native 1a2p<br>all but 1a2p |
|---|---|---|
| **PARAM19** | | |
| Average | 3.3 | 3.4 |
| SD | 3.5 | 3.6 |
| Max | 46.3 | 49.0 |
| | | |
| **PARAM22** | | |
| Average | 3.2 | 3.4 |
| SD | 3.4 | 3.4 |
| Max | 43.9 | 43.8 |

Table 1: Cross validation of FACTS. The values are in kcal/mol and represent atomic electrostatic solvation energy deviations from fdP data calculated with unit charges and $\varepsilon_m = 1$. In the second column, the training set for FACTS parameter optimization is identical to the test set and consists of 81 and 72 protein structures for PARAM19 and PARAM22, respectively. In the third column, training and test sets are disjunct; the training set consists of only the native structure of barnase (1a2p) while the test set consists of all the remaining structures.

| | FACTS<br>$\kappa = 4$ | FACTS<br>$\kappa = 8$ | FACTS<br>$\kappa = 12$ | GBMV2<br>$\kappa = 8$ | GBMVgrid<br>$\kappa = 8$ |
|---|---|---|---|---|---|
| **PARAM19** | | | | | |
| Average [%] | 2.12 | 1.43 | 1.36 | 2.37 | 2.04 |
| SD [%] | 1.64 | 1.26 | 1.23 | 1.34 | 1.32 |
| Max [%] | 10.09 | 10.90 | 11.35 | 8.62 | 8.76 |
| | | | | | |
| **PARAM22** | | | | | |
| Average [%] | 3.54 | 1.90 | 1.56 | 1.29 | 2.29 |
| SD [%] | 1.96 | 1.27 | 1.14 | 1.03 | 1.48 |
| Max [%] | 12.42 | 7.65 | 7.22 | 6.26 | 7.21 |

Table 2: Percentage error of electrostatic solvation energy values of 2900 protein conformations (100 conformations from each of 29 trajectories). The parameter $\kappa$ is in the interaction term of equation (13). The benchmark are the fdP values with $\varepsilon_m = 1$.

| | FACTS<br>$\kappa = 4$ | FACTS<br>$\kappa = 8$ | FACTS<br>$\kappa = 12$ | GBMV2<br>$\kappa = 8$ | GBMVgrid<br>$\kappa = 8$ |
|---|---|---|---|---|---|
| **PARAM19** | | | | | |
| Average | 26.36 | 25.12 | 24.90 | 20.05 | 14.43 |
| SD | 21.76 | 21.06 | 20.84 | 15.47 | 9.59 |
| Max | 94.90 | 91.10 | 89.60 | 59.10 | 38.10 |
| | | | | | |
| **PARAM22** | | | | | |
| Average | 17.41 | 16.05 | 15.95 | 13.95 | 10.39 |
| SD | 9.54 | 8.82 | 8.72 | 8.31 | 4.96 |
| Max | 37.10 | 34.80 | 34.70 | 31.30 | 21.80 |

Table 3: Differences in electrostatic solvation energy from *pairs* of protein conformations ($\Delta\Delta G$). A total of 29 x 4950 values of solvation energy differences were calculated. All values are in kcal/mol. The parameter $\kappa$ is in the interaction term of equation (13). The benchmark are the fdP values with $\varepsilon_m = 1$.

| PDB | residues | $\langle\rangle_{10}$ | $\langle\rangle_{20}$ | $\langle\rangle_{30}$ | $\langle\rangle_{40}$ | $\langle\rangle_{50}$ | $\langle\rangle_{60}$ | $\langle\rangle_{70}$ | $\langle\rangle_{80}$ | $\langle\rangle_{90}$ | $\langle\rangle_{100}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1cb3 | 11 | 2.9 | 3.2 | 3.8 | 3.7 | 3.8 | 3.9 | 4.0 | 4.0 | 4.0 | 4.0 |
| 1l2y | 20 | 1.1 | 1.0 | 1.1 | 1.1 | 1.4 | 2.7 | 2.4 | 2.1 | 2.0 | 1.8 |
| Beta3s[a] | 20 | 2.1 | 2.1 | 2.5 | 2.5 | 2.6 | 2.5 | 2.3 | 2.6 | 2.4 | 2.5 |
| 1f8a | 33 | 1.2 | 1.1 | 1.0 | 1.1 | 1.3 | 1.5 | 1.2 | 1.4 | 2.8 | 3.5 |
| 1vii | 36 | 2.1 | 2.0 | 2.0 | 2.4 | 2.7 | 1.9 | 1.9 | 1.9 | 2.0 | 2.0 |
| 1abz | 38 | 3.9 | 4.5 | 5.6 | 5.7 | 5.6 | 5.8 | 5.7 | 5.8 | 5.7 | 5.7 |
| 1crn | 46 | 1.2 | 1.1 | 1.0 | 0.9 | 0.7 | 0.8 | 0.9 | 0.8 | 0.8 | 1.0 |
| 1enh | 54 | 1.6 | 1.4 | 2.0 | 1.9 | 1.4 | 3.1 | 3.7 | 3.7 | 3.7 | 3.8 |
| 1pgb | 56 | 1.0 | 1.0 | 1.2 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 |
| 1bpi | 58 | 1.8 | 2.0 | 1.9 | 1.9 | 1.9 | 1.9 | 1.9 | 2.0 | 2.1 | 2.3 |
| 1fmk | 59 | 1.4 | 1.4 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 |
| 2ci2 | 65 | 1.6 | 1.7 | 1.7 | 1.8 | 1.9 | 2.2 | 2.3 | 2.3 | 2.2 | 2.1 |
| 2a3d | 73 | 3.6 | 3.7 | 3.8 | 4.0 | 3.6 | 3.6 | 3.7 | 3.7 | 3.7 | 3.7 |
| 1ubq | 76 | 1.5 | 1.4 | 1.6 | 1.9 | 1.9 | 1.9 | 1.9 | 1.9 | 1.9 | 1.9 |
| 1pht | 83 | 2.0 | 1.8 | 2.0 | 2.2 | 2.1 | 2.2 | 2.2 | 2.1 | 2.2 | 2.2 |
| 1hdn | 85 | 2.0 | 2.1 | 1.9 | 1.9 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 |
| 1a2p | 108 | 2.4 | 3.4 | 3.6 | 3.5 | 3.6 | 3.4 | 3.5 | 3.6 | 3.4 | 3.5 |
| | | | | | | | | | | | |
| 1cb3 | 11 | 3.6 | 3.6 | 3.9 | 3.2 | 3.5 | 3.5 | 3.8 | 4.0 | 4.0 | 3.9 |
| 1l2y | 20 | 1.6 | 1.1 | 1.1 | 1.2 | 1.3 | 1.1 | 1.1 | 1.3 | 1.1 | 1.1 |
| Beta3s[a] | 20 | 3.3 | 3.7 | 3.7 | 3.5 | 3.2 | 3.2 | 3.2 | 3.3 | 3.3 | 3.2 |
| 1f8a | 33 | 1.2 | 1.7 | 1.1 | 1.2 | 1.1 | 1.2 | 1.3 | 1.4 | 1.4 | 1.3 |
| 1vii | 36 | 2.3 | 2.1 | 2.1 | 2.0 | 1.9 | 1.9 | 2.0 | 1.9 | 2.0 | 1.9 |
| 1abz | 38 | 3.9 | 3.8 | 4.4 | 4.1 | 4.0 | 4.2 | 4.1 | 4.0 | 4.1 | 4.3 |
| 1crn | 46 | 0.9 | 0.9 | 0.9 | 0.9 | 1.1 | 0.9 | 0.9 | 1.0 | 0.9 | 1.1 |
| 1enh | 54 | 1.1 | 1.1 | 1.3 | 1.6 | 2.2 | 2.3 | 3.7 | 3.4 | 3.1 | 3.0 |
| 1pgb | 56 | 0.9 | 0.9 | 1.0 | 1.0 | 0.9 | 1.0 | 1.1 | 1.0 | 1.1 | 1.2 |
| 1bpi | 58 | 2.0 | 1.9 | 1.9 | 1.9 | 2.0 | 2.3 | 2.2 | 2.2 | 2.2 | 2.1 |
| 1fmk | 59 | 1.6 | 1.7 | 1.7 | 1.8 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.2 |
| 2ci2 | 65 | 1.4 | 1.5 | 2.1 | 2.0 | 2.1 | 2.2 | 2.2 | 2.2 | 2.1 | 2.2 |
| 2a3d | 73 | 3.4 | 3.5 | 3.4 | 3.4 | 3.4 | 3.3 | 3.4 | 3.3 | 3.2 | 3.2 |
| 1ubq | 76 | 1.8 | 1.8 | 1.8 | 1.8 | 1.9 | 2.0 | 2.1 | 2.0 | 2.0 | 2.1 |
| 1pht | 83 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.1 | 2.0 | 2.1 | 2.2 | 2.0 |
| 1hdn | 85 | 1.4 | 1.5 | 1.5 | 1.5 | 1.5 | 1.6 | 1.7 | 1.6 | 1.5 | 1.6 |
| 1a2p | 108 | 2.4 | 1.9 | 1.9 | 2.0 | 2.1 | 2.1 | 2.0 | 2.2 | 2.2 | 2.0 |

Table 4: Deviation from the native structure during MD simulations at 300 K. Individual columns contain values in Å of the $C_\alpha$-RMSD from the native structure averaged over 10 ns intervals, e.g., for the last column $<>_{100}$ the $C_\alpha$-RMSD was averaged over the 90-100 ns interval. The simulations were performed with FACTS PARAM22, $\kappa = 4$, $\varepsilon_m = 1.0$, and using $\gamma = 0.015$ kcal mol$^{-1}$ Å$^{-2}$ (top) and $\gamma = 0.025$ kcal mol$^{-1}$ Å$^{-2}$ (bottom). [a]Beta3s is a three-stranded antiparallel $\beta$-sheet peptide [40, 52].

| | | PARAM19 | | PARAM22 | |
|---|---|---|---|---|---|
| PDB | residues | atoms | CPU-days | atoms | CPU-days |
| 1cb3 | 11 | 112 | 0.6 | 186 | 3.5 |
| Beta3s[a] | 20 | 215 | 2.2 | 329 | 8.4 |
| 1crn | 46 | 396 | 3.8 | 642 | 21.9 |
| 2ci2 | 65 | 636 | 6.6 | 1076 | 43.1 |
| 1a2p | 108 | 1073 | 14.5 | 1700 | 71.3 |

Table 5: Computation time required for a 100-ns MD simulations with FACTS. All simulations were performed on a single Opteron 1.8 GHz processor. [a]Beta3s is a three-stranded antiparallel $\beta$-sheet peptide [40, 52].

# FACTS: Fast Analytical Continuum Treatment of Solvation

## Supplementary Material

Urs Haberthür and Amedeo Caflisch*

Department of Biochemistry, University of Zurich

Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

E-mail: caflisch@bioc.unizh.ch

Phone: +41 44 635 55 21

Fax: +41 44 635 68 62

October 16, 2006

*Corresponding author

# 1 Results with interior dielectric $\varepsilon_m = 2$

## 1.1 Dependency of FACTS Parameters on Training Set of Proteins

Two different setups are chosen to assess the dependency of the FACTS solvation parameters on the training set. In the first one, training and test sets are identical and consist of all proteins. In the second one, training and test sets are disjunct. The training set consists only of the native state of 1a2p, and the test set of all proteins without the native state of 1a2p. The results are shown in Table 1 and Figures 1 and 2. They demonstrate that the dependency of the FACTS parameters on the training set is marginal. The FACTS parameters for PARAM19 and PARAM22 for $\varepsilon_m = 1$ and $\varepsilon_m = 2$ are given in tables 3, 4, 5, 6, 7, 8, 9, and 10, respectively.

| | $\varepsilon_m = 1$ | | $\varepsilon_m = 2$ | |
|---|---|---|---|---|
| training set | all | native 1a2p | all | native 1a2p |
| test set | all | all but 1a2p | all | all but 1a2p |
| **PARAM19** | | | | |
| aver | 3.322 | 3.400 | 1.428 | 1.496 |
| sig | 3.486 | 3.554 | 1.545 | 1.555 |
| max | 46.255 | 49.042 | 18.119 | 18.040 |
| **PARAM22** | | | | |
| aver | 3.200 | 3.353 | 1.494 | 1.577 |
| sig | 3.370 | 3.442 | 1.569 | 1.600 |
| max | 43.860 | 43.765 | 20.935 | 21.267 |

Table 1: Cross validation of FACTS. The values are in kcal/mol and represent atomic solvation energy deviations from fdP data calculated with unit charges. In the second and fourth columns, the training set for FACTS parameter optimization is identical to the test set and consists of 81 and 72 protein structures for PARAM19 and PARAM22, respectively. In the third and fifth columns, training and test sets are disjunct; the training set consists of only the native structure of barnase (1a2p) while the test set consists of all the remaining structures.

## 1.2 Atomic (or Self) Electrostatic Solvation Energy

Figure 3 shows atomic solvation energy values calculated by FACTS, GBMV2, and GBMVgrid versus the benchmark fdP values with $\varepsilon_m = 2$. The numerical approach GBMVgrid is the most accurate method, followed by GBMV2 and FACTS. However, the maximal absolute error is largest for GBMV2 because of some significant outliers, yet its statistical spread is between the one of FACTS and GBMVgrid. This is observed for both PARAM19 and PARAM22.

# 2 Atomic SASA

Figure 4 compares SASA values calculated by FACTS and the Hasel formula [1].

# 3 Pairwise Electrostatic Energies and Their Sums

The results for $\epsilon_m = 1$ and $\epsilon_m = 2$ are given in Figures 5, 6, 7, and 8.

## 4 Electrostatic Solvation Energy of Protein Conformations

See Tables 11-15.

## 5 Energy in Solution

The electrostatic free energy of solvation $\Delta G$ can be written as the sum of a self-energy term and an interaction energy term

$$\Delta G = \sum_i \Delta G_i + \sum_{i<j} \left( G_{ij}^{sl} - G_{ij}^{vac} \right) \tag{1}$$

where the superscript *sl* indicates the solvent. The electrostatic free energy in solution $G$ can be formally written as the sum of electrostatic free energy in vacuo $G^{vac}$ and $\Delta G$ [2]

$$G = \Delta G + G^{vac} \tag{2}$$

Assuming that the solute (macromolecule) has the same dielectric constant as vacuo (i.e., $\varepsilon_m = 1$), one has a system with homogeneous dielectric response where Born's self energy formula and Coulomb's law apply, so that

$$G^{vac} = \sum_i \frac{q_i^2}{2\varepsilon_m r_i^{vdW}} + \sum_{i<j} \frac{q_i q_j}{\varepsilon_m r_{ij}} \tag{3}$$

Note that $\varepsilon_m$ is kept in the above equation because the homogeneous dielectric response is present for a solute in any environment with $\varepsilon_{out} = \varepsilon_m$. The value of $\varepsilon_m = 1$ is usually adopted to be consistent with the assumptions under which the partial charge of common force fields have been derived [3]. Combining equations (1), (2), and (3), the electrostatic free energy in solution can be written as

$$G = \sum_i \Delta G_i + \sum_i \frac{q_i^2}{2\varepsilon_m r_i^{vdW}} + \sum_{i<j} G_{ij}^{sl} \tag{4}$$

It is important to note that the Born term $\sum_i \frac{q_i^2}{2\varepsilon_m r_i^{vdW}}$ is an additive constant because it does not depend on the solute configuration.

It has been shown previously that a high correlation between approximated and exact solvation energies does not necessarily imply a good correlation between approximated and exact energies in solution [4, 5]. This is because vacuo pair interaction energies can always be calculated exactly and they dominate the correlation between approximated and exact solvation energies. Eliminating vacuo energies from the comparison, i.e., comparing energies in solution instead of solvation energies, is therefore another useful test of the accuracy of a solvation model [4, 5].

For this purpose high temperature unfolding simulations at 450 K for 50 ns using an implicit solvation model [6] of 29 proteins were performed. Coordinates were saved every 10 ps and all snapshots were sorted according to

increasing radius of gyration (RG). A total of 100 conformations were chosen from each trajectory as follows: every 20th conformation from the 500 snapshots with the lowest RG (25 conformations), every 20th conformation from the 500 snapshots with the largest RG (25 conformations), and every 80th conformation from the remaining 4000 snapshots (50 conformations). The 100 conformations of each protein cover a wide range of RMSD and RG. For each structure in every trajectory the quantity $G$ was calculated according to equations (4), for FACTS, GBMV2, and GBMVgrid, and compared to the fdP values. Note that the Born term $\sum_i \frac{q_i^2}{2\varepsilon_m r^{vdW}}$ in equation (4) is neglected since it is an additive constant as mentioned above. The results are shown in Figure 9 and Tables 16-18.

## 6 Molecular Dynamics Results with FACTS PARAM19

Table 2 shows the results of 100-ns molecular dynamics runs with FACTS PARAM19 (i.e., polar hydrogen CHARMM force field [3]).

| PDB | residues | $\langle\rangle_{10}$ | $\langle\rangle_{30}$ | $\langle\rangle_{40}$ | $\langle\rangle_{50}$ | $\langle\rangle_{60}$ | $\langle\rangle_{70}$ | $\langle\rangle_{80}$ | $\langle\rangle_{90}$ | $\langle\rangle_{100}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1cb3 | 11 | 3.3 | 3.7 | 3.4 | 3.5 | 3.4 | 3.6 | 3.5 | 4.5 | 4.8 |
| Beta3s[a] | 20 | 2.1 | 2.2 | 2.2 | 2.2 | 2.2 | 2.2 | 2.2 | 2.1 | 2.4 |
| 1crn | 46 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 |
| 2ci2 | 65 | 1.9 | 2.1 | 2.1 | 2.0 | 2.1 | 2.0 | 2.1 | 2.0 | 2.0 |
| 1a2p | 108 | 3.4 | 4.3 | 4.5 | 4.5 | 4.5 | 4.6 | 4.5 | 4.4 | 4.4 |

Table 2: Deviation from the native structure during molecular dynamics simulations at 300 K. Individual columns contain values of the $C_\alpha$-RMSD from the native structure averaged over 10 ns intervals, e.g., for the last column $<>_{100}$ the $C_\alpha$-RMSD was averaged over the 90-100 ns interval. The simulations were performed with FACTS PARAM19, $\kappa = 4$, $\varepsilon_m = 1.0$, and $\gamma = 0.025$ kcal mol$^{-1}$ Å$^{-2}$. [a]Beta3s is a three-stranded antiparallel $\beta$-sheet peptide [7, 8].

## References


[1] Hasel, W.; Hendrickson, T. F. and Still, W. C., *Tetrahedron Comput. Methodol.*, 1988, **1**, 103–116.

[2] Schaefer, M. and Karplus, M., *J. Phys. Chem.*, 1996, **100**, 1578–1599.

[3] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S. and Karplus, M., *J. Comput. Chem.*, 1983, **4**, 187–217.

[4] Scarsi, M. and Caflisch, A., *J. Comput. Chem.*, 1999, **14**, 1533–1536.

[5] Bashford, D. and Case, D. A., *Annu. Rev. Phys. Chem.*, 2000, **51**, 129–152.

[6] Ferrara, P.; Apostolakis, J. and Caflisch, A., *Proteins: Structure, Function, and Bioinformatics*, 2002, **46**, 24–33.

[7] Ferrara, P. and Caflisch, A., *Proc. Natl. Acad. Sci. USA.*, 2000, **97**, 10780–10785.

[8] De Alba, E.; Santoro, J.; Rico, M. and Jiménez, M. A., *Protein Science*, 1999, **8**, 854–865.

[9] Lee, M. S.; Feig, M.; Salsbury, F. R. and Brooks III, C. L., *J. Comput. Chem.*, 2003, **24**, 1348–1356.

[10] Lee, M. S.; Salsbury, F. R. and Brooks III, C. L., *J. Chem. Phys.*, 2002, **116**, 10606–10614.

[11] Lee, B. and Richards, F. M., *J. Mol. Biol.*, 1971, **55**, 379–400.

## 7 Supplementary Figures

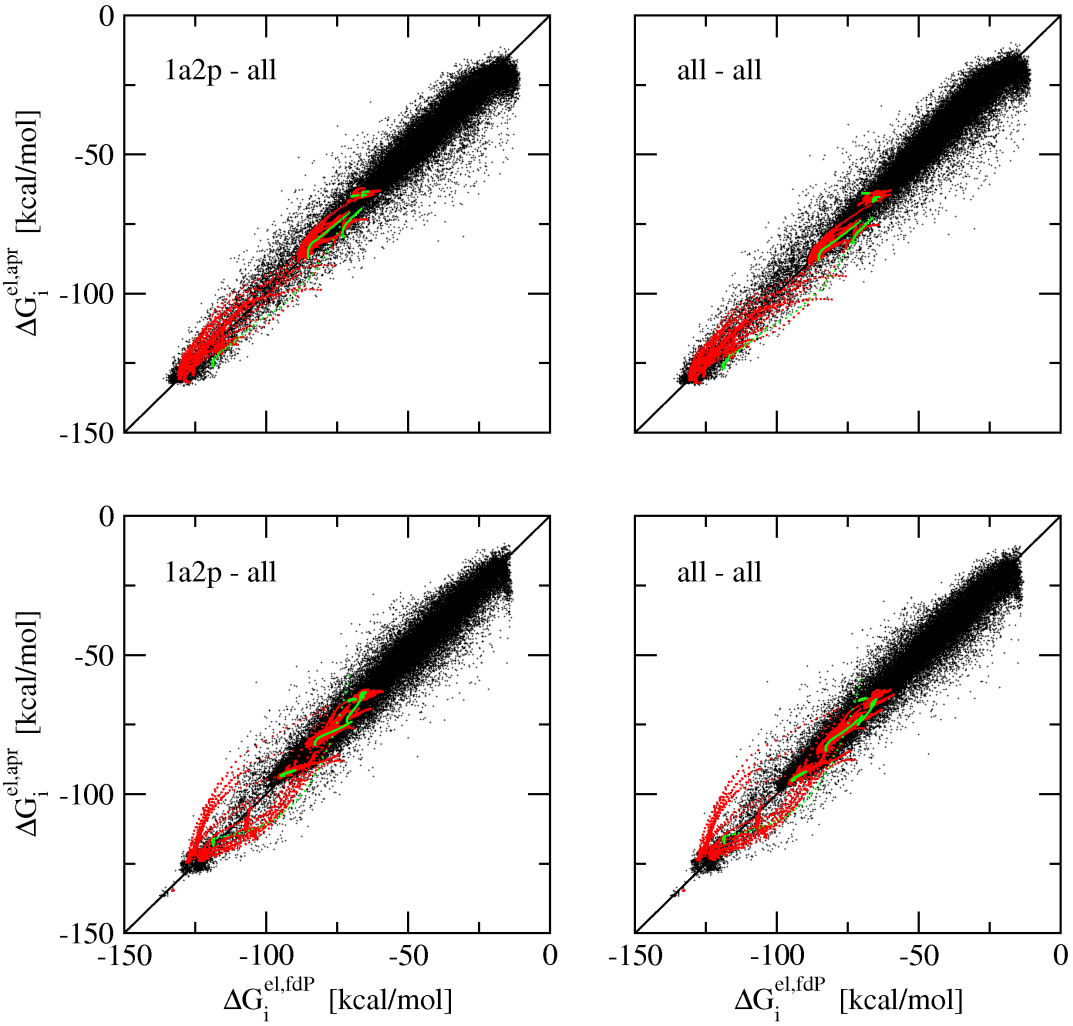


Figure 1: Cross validation of the FACTS solvation parameters ($\varepsilon_m = 1$). On the abscissa the exact values are given, and on the ordinate the approximated values. In the left column, training and test set are disjunct. The training set consists only of the native state of 1a2p and the test set of all protein structures but the native state of 1a2p, pairs of ionic side chains, and the conformations of the N-methyl-acetamide dimer. In the right column, the training set consists of all protein structures and the test set of all protein structures, pairs of ionic side chains, and the conformations of the N-methyl-acetamide dimer. Unit charges are used because they allow for a more stringent comparison that is not affected by the charge parameter set. The protein conformations are in black, the pairs of ionic side chains in red, and the conformations of the N-methyl-acetamide dimer in green. There is essentially no difference between the results obtained with the two parameter sets. (Top): Atomic solvation energy values of 77'609 atoms from 1'082 structures from 37 molecular systems calculated with the van der Waals radii from PARAM19. (Bottom) Atomic solvation energy values of 90'747 atoms from 1'073 structures from 32 molecular systems calculated with the van der Waals radii from PARAM22.

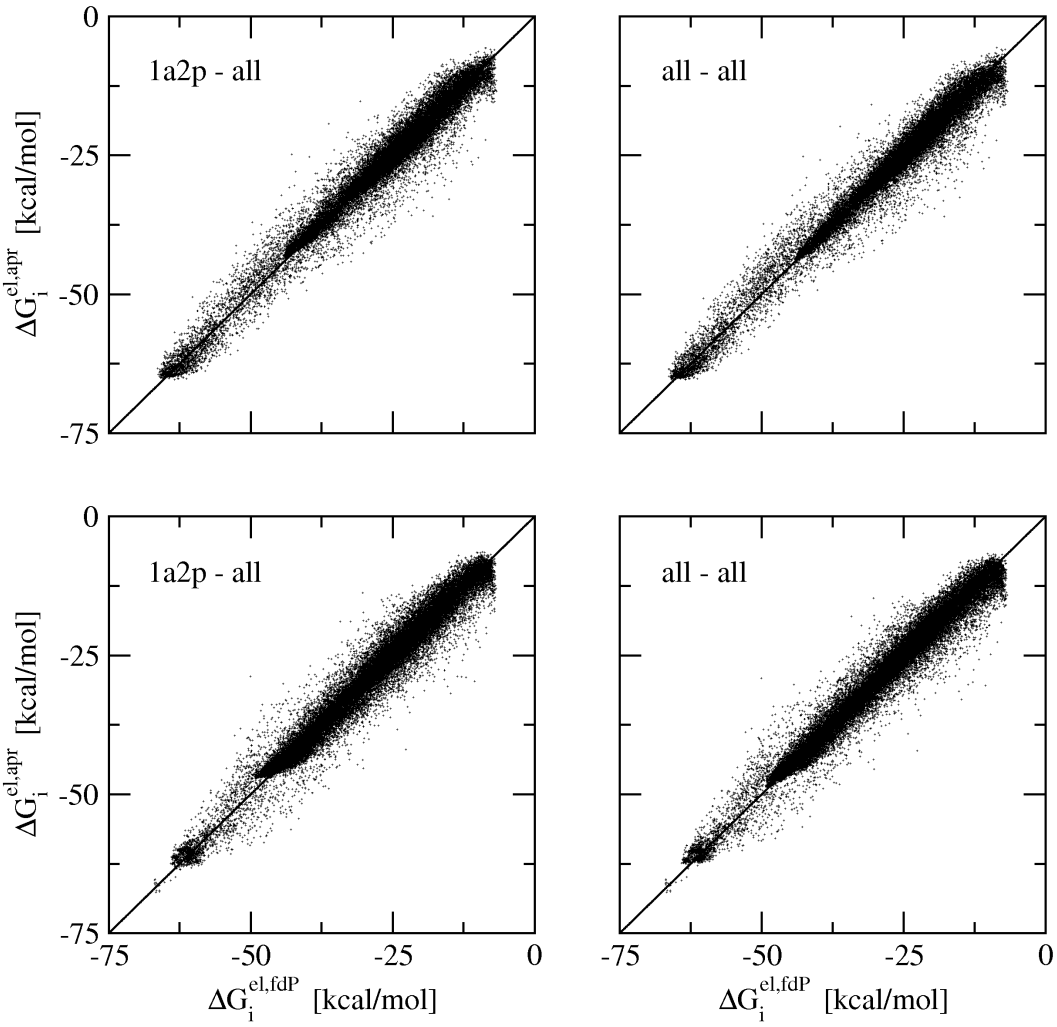


Figure 2: Cross validation of the FACTS solvation parameters ($\varepsilon_m = 2$). See legend of Figure 1 for details.

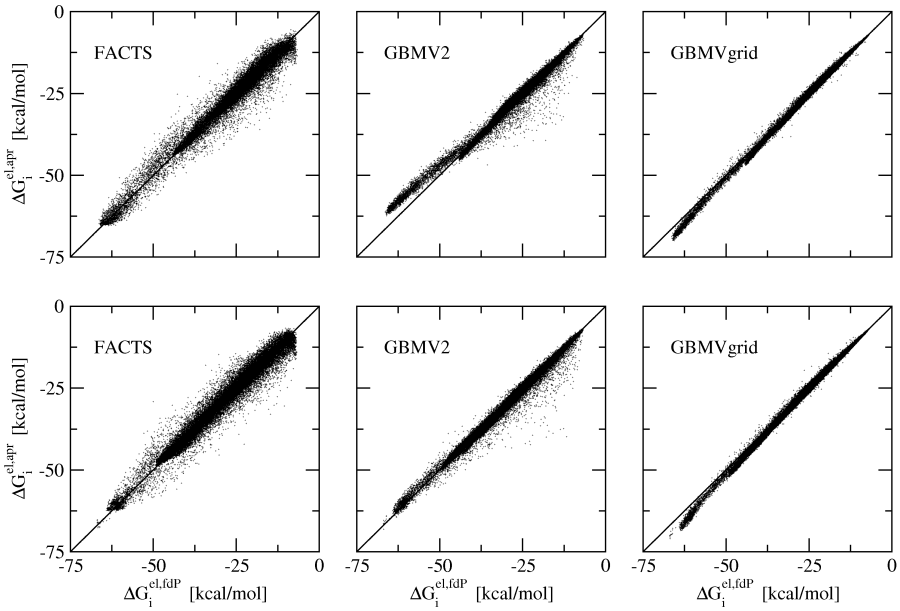


Figure 3: Comparison between FACTS and GBMV. The plot shows the atomic electrostatic solvation energy evaluated with unit charges and $\varepsilon_m = 2$ for PARAM19 (top) and PARAM22 (bottom). Unit charges are used because they allow for a more stringent comparison that is not affected by the charge parameter set. On the abscissa the fdP derived values are given, and on the ordinate the approximated values. (Top) Slope, correlation, and maximal absolute error for the 31'036 atoms from 63 protein structures are 0.970, 0.984, and 18.1 kcal/mol for FACTS; 0.919, 0.991, and 21.3 kcal/mol for GBMV2 [9]; 1.042, 0.998, and 7.8 kcal/mol for GBMVgrid [10]. (Bottom) Slope, correlation, and maximal absolute error for the 38'514 atoms from 57 structures are 0.967, 0.983, and 20.9 kcal/mol for FACTS; 0.977, 0.994, and 28.2 kcal/mol for GBMV2; 1.053, 0.998, and 6.9 kcal/mol for GBMVgrid.

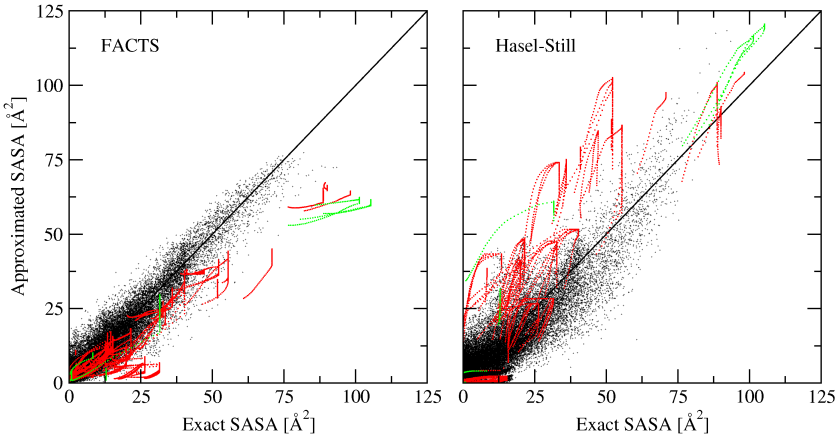


Figure 4: Atomic SASA of 1'082 structures with PARAM19 and $\varepsilon_m = 1$. In the left plot the data are obtained with the FACTS model, and in the right plot with the approximated formula by Hasel et al. [1]. The benchmark are the exact values of atomic SASA [11]. The color coding is the same as in Figure 1.

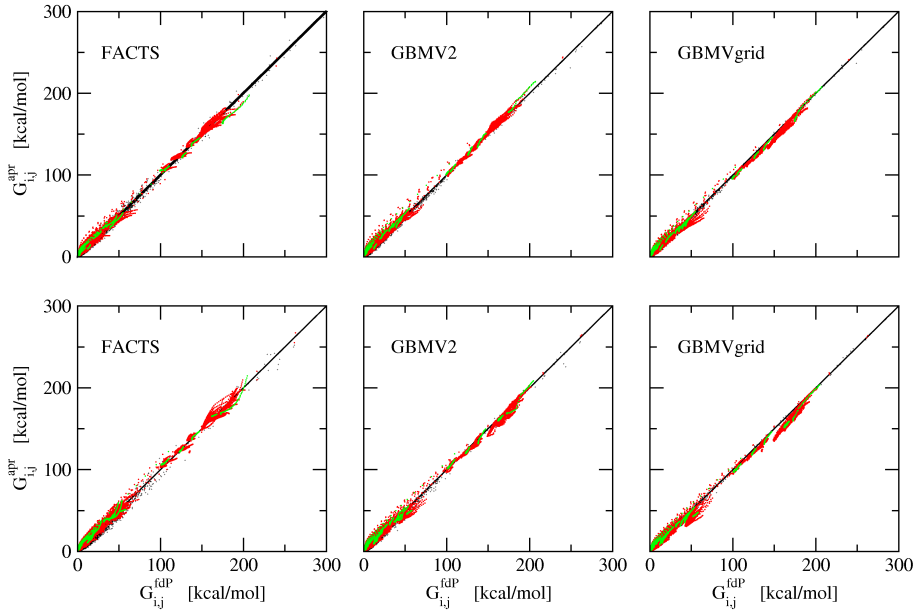


Figure 5: Interaction energy values for PARAM19 (top) and PARAM22 (bottom), $\varepsilon_m = 1$, and unit charges for all atoms. On the abscissa the fdP derived values are given, and on the ordinate the approximated values. The color coding is the same as in Figure 1.

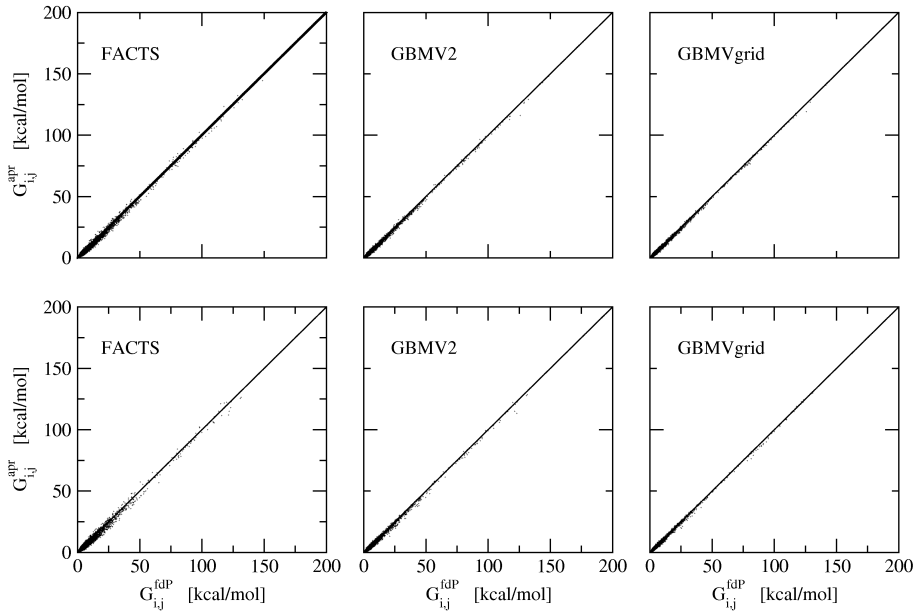


Figure 6: Interaction energy values for PARAM19 (top) and PARAM22 (bottom), $\varepsilon_m = 2$, and unit charges for all atoms. On the abscissa the fdP derived values are given, and on the ordinate the approximated values.

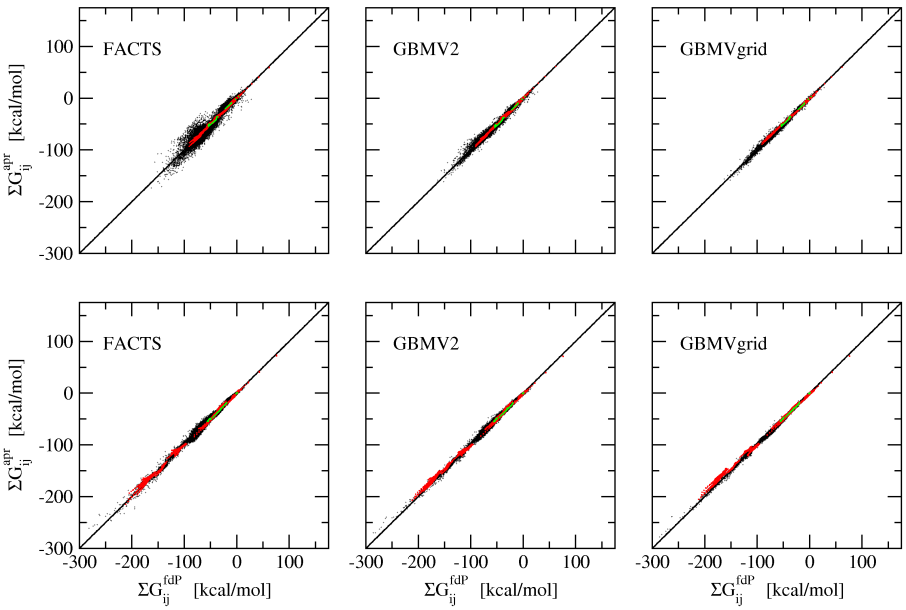


Figure 7: Comparison of sums of interaction energy values between FACTS and GBMV. For each solute atom the sum over all its interaction energies, using partial charges, is calculated with $\varepsilon_m = 1$. On the abscissa the fdP derived values are given, and on the ordinate the approximated values. (Top) PARAM19. (Bottom) PARAM22.

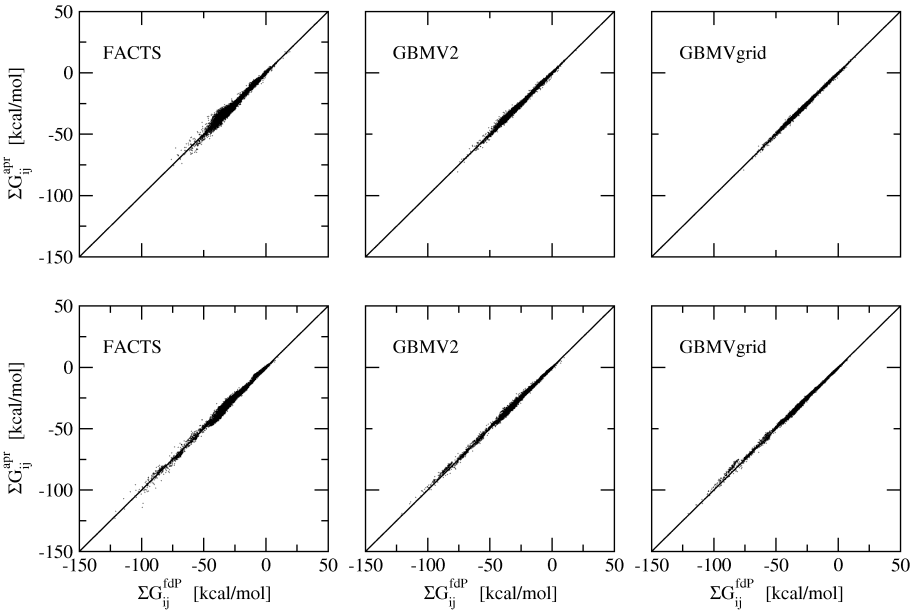


Figure 8: Comparison of sums of interaction energy values between FACTS and GBMV. For each solute atom the sum over all its interaction energies, using partial charges, is calculated with $\varepsilon_m = 2$. On the abscissa the fdP derived values are given, and on the ordinate the approximated values. (Top) PARAM19. (Bottom) PARAM22.

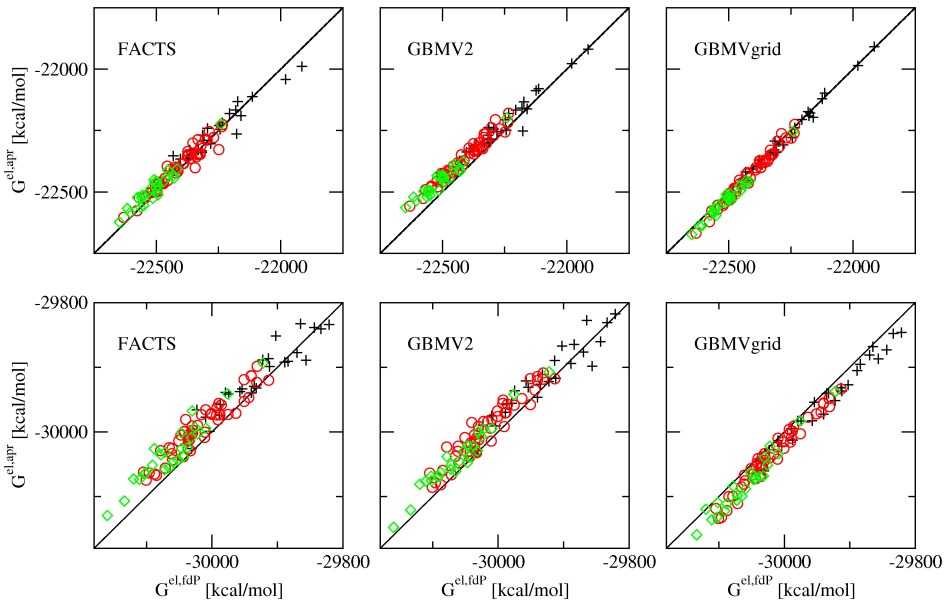


Figure 9: Comparison of energy in solution values between FACTS and GBMV. (The energy in solution of a conformation is its solvation energy plus the vacuo pair interaction energies.) The values for 100 conformations of 1a2p are shown for PARAM19 (top) and PARAM22 (bottom) with $\varepsilon_m = 1$. The structures are chosen along a high temperature unfolding trajectory. Different symbols discriminate between different ranges of the radius of gyration. Pluses and diamonds represent the 25 conformations with small and large radius of gyration, respectively, and circles the 50 intermediate ones. On the abscissa the fdP derived values are given, and on the ordinate the approximated values.

# 8 FACTS Parameters

| $r_{vdW}$ | atom type | $b_1$ | $b_2$ | $a_2$ | $a_3$ | $R_{sphere}$ |
|---|---|---|---|---|---|---|
| 1.0 | H | -241 | -1.169 | 0.00355 | 167 | 7.6 |
| 1.6 | N, O | -307 | -0.966 | 0.00223 | 577 | 8.9 |
| 1.89 | S | 278 | -1.008 | 0.00196 | 1079 | 10.0 |
| 2.1 | C,CR1E | -44 | -0.830 | 0.00179 | 1109 | 10.0 |
| 2.165 | CH3E | 392 | -0.861 | 0.00265 | 1202 | 10.0 |
| 2.235 | CH2E | 230 | -0.965 | 0.00217 | 1111 | 9.8 |
| 2.365 | CH1E | 738 | -1.167 | 0.00208 | 1259 | 10.0 |

Table 3: The 35 FACTS electrostatic solvation parameters for PARAM19 and $\varepsilon_m = 1$.

| $r_{vdW}$ | atom type | $d_1$ | $d_2$ | $c_2$ | $c_3$ |
|---|---|---|---|---|---|
| 1.0 | H | -5310 | -2.836 | 0.00095 | -6343 |
| 1.6 | N, O | -7217 | -5.880 | 0.00072 | -8337 |
| 1.89 | S | -449 | -1.848 | 0.00311 | -800 |
| 2.1 | C,CR1E | -2930 | -2.462 | 0.00168 | -2935 |
| 2.165 | CH3E | -626 | -1.426 | 0.00248 | -716 |
| 2.235 | CH2E | -330 | -1.561 | 0.00318 | -526 |
| 2.365 | CH1E | 291 | -2.109 | 0.00371 | -433 |

Table 4: The 28 FACTS surface parameters for PARAM19 and $\varepsilon_m = 1$.

| $r_{vdW}$ | atom type | $b_1$ | $b_2$ | $a_2$ | $a_3$ | $R_{sphere}$ |
|---|---|---|---|---|---|---|
| 1.0 | H | -318 | -1.124 | 0.00339 | 116 | 7.4 |
| 1.6 | N, O | -393 | -0.816 | 0.00226 | 518 | 8.5 |
| 1.89 | S | 243 | -0.955 | 0.00254 | 862 | 9.2 |
| 2.1 | C,CR1E | -263 | -0.573 | 0.00184 | 1012 | 9.6 |
| 2.165 | CH3E | 194 | -0.749 | 0.00236 | 1183 | 10.0 |
| 2.235 | CH2E | 47 | -0.828 | 0.00222 | 1000 | 9.4 |
| 2.365 | CH1E | 439 | -0.945 | 0.00186 | 1265 | 10.0 |

Table 5: The 35 FACTS electrostatic solvation parameters for PARAM19 and $\varepsilon_m = 2$.

| $r_{vdW}$ | atom type | $d_1$ | $d_2$ | $c_2$ | $c_3$ |
|---|---|---|---|---|---|
| 1.0 | H | -1461 | -1.525 | 0.00279 | -1831 |
| 1.6 | N, O | -5955 | -5.205 | 0.00088 | -6818 |
| 1.89 | S | -288 | -1.853 | 0.00423 | -566 |
| 2.1 | C,CR1E | -2017 | -2.293 | 0.00222 | -2075 |
| 2.165 | CH3E | -626 | -1.426 | 0.00248 | -716 |
| 2.235 | CH2E | -153 | -1.585 | 0.00395 | -356 |
| 2.365 | CH1E | 291 | -2.109 | 0.00371 | -433 |

Table 6: The 28 FACTS surface parameters for PARAM19 and $\varepsilon_m = 2$.

| $r_{vdW}$ | atom type | $b_1$ | $b_2$ | $a_2$ | $a_3$ | $R_{sphere}$ |
|---|---|---|---|---|---|---|
| 0.2245 | H | -14 | -1.910 | 0.00234 | -807 | 7.2 |
| 0.9 | H | 74 | -1.348 | 0.00561 | 227 | 7.1 |
| 1.32 | H | -77 | -1.334 | 0.00267 | 353 | 8.2 |
| 1.3582 | H | -140 | -1.454 | 0.00250 | 403 | 8.4 |
| 1.468 | H | 70 | -0.585 | 0.00409 | 640 | 8.0 |
| 1.7 | O | -250 | -0.949 | 0.00231 | 674 | 8.6 |
| 1.77 | O | -234 | -0.827 | 0.00244 | 696 | 8.5 |
| 1.8 | C | -500 | -0.521 | 0.00215 | 628 | 8.3 |
| 1.85 | N | -477 | -0.513 | 0.00207 | 760 | 8.5 |
| 1.975 | S | 119 | -0.717 | 0.00253 | 959 | 8.9 |
| 1.9924 | C | -95 | -0.979 | 0.00189 | 944 | 9.2 |
| 2.0 | C | 32 | -0.698 | 0.00202 | 950 | 8.8 |
| 2.06 | C | 269 | -1.034 | 0.00189 | 1267 | 9.9 |
| 2.175 | C | 63 | -0.941 | 0.00225 | 980 | 8.9 |
| 2.275 | C | 673 | -1.239 | 0.00185 | 1243 | 9.5 |

Table 7: The 75 FACTS electrostatic solvation parameters for PARAM22 and $\varepsilon_m = 1$.

| $r_{vdW}$ | atom type | $d_1$ | $d_2$ | $c_2$ | $c_3$ |
|---|---|---|---|---|---|
| 0.2245 | H | -11122 | 49.835 | 5.32615 | -4670 |
| 0.9 | H | -1890 | -4.764 | 0.00208 | -3145 |
| 1.32 | H | -22733 | -12.329 | 0.00026 | -27557 |
| 1.3582 | H | -33836 | -27.374 | 0.00016 | -41991 |
| 1.468 | H | -346 | -1.571 | 0.00524 | -488 |
| 1.7 | O | -428 | -1.616 | 0.00384 | -598 |
| 1.77 | O | -1000 | -1.446 | 0.00276 | -1050 |
| 1.8 | C | -15278 | -7.460 | 0.00049 | -14322 |
| 1.85 | N | -369 | -1.607 | 0.00453 | -479 |
| 1.975 | S | -1977 | -3.375 | 0.00184 | -2466 |
| 1.9924 | C | -2630 | -0.924 | 0.00158 | -2519 |
| 2.0 | C | -17736 | -2.008 | 0.00057 | -13818 |
| 2.06 | C | -17 | -1.741 | 0.00307 | -568 |
| 2.175 | C | 14 | -1.997 | 0.00461 | -407 |
| 2.275 | C | 457 | -2.444 | 0.00434 | -445 |

Table 8: The 60 FACTS surface parameters for PARAM22 and $\varepsilon_m = 1$.

| $r_{vdW}$ | atom type | $b_1$ | $b_2$ | $a_2$ | $a_3$ | $R_{sphere}$ |
|---|---|---|---|---|---|---|
| 0.2245 | H | -53 | -1.920 | 0.00232 | -848 | 7.0 |
| 0.9 | H | 55 | -1.706 | 0.00710 | 114 | 6.8 |
| 1.32 | H | -107 | -1.323 | 0.00257 | 342 | 8.3 |
| 1.3582 | H | -151 | -1.343 | 0.00255 | 405 | 8.3 |
| 1.468 | H | -28 | -0.793 | 0.00406 | 483 | 7.5 |
| 1.7 | O | -297 | -0.914 | 0.00227 | 642 | 8.5 |
| 1.77 | O | -357 | -0.727 | 0.00224 | 677 | 8.5 |
| 1.8 | C | -519 | -0.387 | 0.00215 | 627 | 8.3 |
| 1.85 | N | -489 | -0.445 | 0.00206 | 746 | 8.4 |
| 1.975 | S | 276 | -0.980 | 0.00224 | 1025 | 9.2 |
| 1.9924 | C | -201 | -0.921 | 0.00179 | 941 | 9.3 |
| 2.0 | C | 4 | -0.634 | 0.00200 | 941 | 8.8 |
| 2.06 | C | 162 | -0.981 | 0.00176 | 1282 | 10.0 |
| 2.175 | C | 9 | -0.868 | 0.00220 | 977 | 8.9 |
| 2.275 | C | 610 | -1.169 | 0.00180 | 1245 | 9.5 |

Table 9: The 75 FACTS electrostatic solvation parameters for PARAM22 and $\varepsilon_m = 2$.

| $r_{vdW}$ | atom type | $d_1$ | $d_2$ | $c_2$ | $c_3$ |
|---|---|---|---|---|---|
| 0.2245 | H | -11122 | 49.835 | 5.32615 | -4670 |
| 0.9 | H | -3375 | -8.253 | 0.00135 | -5403 |
| 1.32 | H | -22733 | -12.329 | 0.00026 | -27557 |
| 1.3582 | H | -9700 | -9.317 | 0.00052 | -12274 |
| 1.468 | H | -1564 | -2.299 | 0.00260 | -1748 |
| 1.7 | O | -428 | -1.616 | 0.00384 | -598 |
| 1.77 | O | -1007 | -1.459 | 0.00278 | -1058 |
| 1.8 | C | -5848 | -4.003 | 0.00117 | -5611 |
| 1.85 | N | -369 | -1.607 | 0.00453 | -479 |
| 1.975 | S | -3864 | -4.757 | 0.00112 | -4664 |
| 1.9924 | C | -228 | -1.281 | 0.00355 | -596 |
| 2.0 | C | -33002 | 29.109 | 0.00028 | -19550 |
| 2.06 | C | -261 | -1.690 | 0.00265 | -770 |
| 2.175 | C | 14 | -1.997 | 0.00461 | -407 |
| 2.275 | C | 457 | -2.444 | 0.00434 | -445 |

Table 10: The 60 FACTS surface parameters for PARAM22 and $\varepsilon_m = 2$.

## 9 Electrostatic Solvation Energy of Protein Conformations

| | | PARAM19 | | | | | PARAM22 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FACTS | FACTS | FACTS | GBMV2 | GBgrid | FACTS | FACTS | FACTS | GBMV2 | GBgrid |
| PDB | residues | $\kappa = 4$ | $\kappa = 8$ | $\kappa = 12$ | $\kappa = 8$ | $\kappa = 8$ | $\kappa = 4$ | $\kappa = 8$ | $\kappa = 12$ | $\kappa = 8$ | $\kappa = 8$ |
| 1cb3 | 11 | 1.577 | 1.702 | 1.750 | 1.380 | 4.589 | 1.686 | 1.988 | 2.224 | 2.694 | 5.476 |
| bet1 | 14 | 0.864 | 1.053 | 1.296 | 0.641 | 4.039 | 1.921 | 0.884 | 0.919 | 1.383 | 3.767 |
| hlx1 | 17 | 0.791 | 0.988 | 1.079 | 1.192 | 4.370 | 1.218 | 1.078 | 1.178 | 2.708 | 5.544 |
| 1l2y | 20 | 1.624 | 0.962 | 0.968 | 2.726 | 3.686 | 2.833 | 1.349 | 1.181 | 1.448 | 4.500 |
| Beta3s[a] | 20 | 1.848 | 0.935 | 1.263 | 3.267 | 3.641 | 3.546 | 1.328 | 1.217 | 0.906 | 3.829 |
| 1f8a | 33 | 1.853 | 1.188 | 1.048 | 3.247 | 1.975 | 2.904 | 1.214 | 0.873 | 0.651 | 2.523 |
| 1abz | 38 | 1.226 | 1.004 | 0.965 | 1.951 | 3.043 | 2.318 | 1.315 | 1.115 | 0.886 | 3.556 |
| 1crn | 46 | 3.995 | 1.276 | 1.183 | 3.645 | 1.762 | 6.327 | 2.214 | 1.335 | 1.384 | 1.966 |
| ins2 | 51 | 2.737 | 1.302 | 1.098 | 1.939 | 2.095 | 4.842 | 2.620 | 2.023 | 0.910 | 1.919 |
| 1enh | 54 | 1.427 | 1.233 | 1.249 | 2.936 | 1.489 | 2.012 | 0.844 | 0.631 | 0.595 | 2.136 |
| 1pgb | 56 | 1.391 | 1.036 | 0.990 | 1.197 | 2.011 | 2.454 | 1.511 | 1.311 | 0.743 | 2.019 |
| 1shg | 57 | 1.555 | 1.339 | 1.300 | 2.501 | 2.041 | 3.225 | 2.064 | 1.778 | 0.723 | 2.942 |
| 1bpi | 58 | 1.749 | 1.383 | 1.326 | 3.033 | 1.124 | 2.312 | 1.188 | 0.940 | 0.720 | 1.810 |
| 1fmk | 59 | 2.688 | 1.369 | 1.238 | 1.924 | 2.015 | 4.813 | 2.456 | 1.852 | 1.628 | 1.397 |
| 2ptl | 61 | 1.783 | 1.165 | 1.104 | 2.074 | 2.189 | 3.571 | 1.987 | 1.610 | 0.872 | 2.692 |
| 2ci2 | 65 | 1.461 | 1.130 | 1.057 | 2.123 | 2.101 | 3.767 | 2.476 | 2.131 | 0.815 | 2.453 |
| 2a3d | 73 | 1.688 | 1.335 | 1.245 | 2.230 | 1.894 | 3.209 | 1.897 | 1.607 | 0.875 | 2.376 |
| 1ubq | 76 | 1.643 | 1.233 | 1.144 | 2.377 | 1.785 | 3.791 | 2.278 | 1.905 | 0.998 | 2.320 |
| 1pht | 83 | 1.598 | 1.141 | 1.047 | 1.322 | 1.591 | 3.279 | 2.217 | 2.000 | 1.200 | 1.129 |
| 1hdn | 85 | 1.740 | 0.956 | 0.852 | 2.078 | 1.430 | 3.883 | 2.225 | 1.783 | 0.819 | 1.909 |
| 1dvd | 98 | 1.695 | 1.236 | 1.130 | 1.952 | 1.433 | 3.681 | 2.386 | 2.038 | 0.931 | 1.967 |
| prph | 104 | 2.992 | 1.692 | 1.456 | 2.594 | 1.209 | 5.399 | 2.932 | 2.271 | 2.686 | 0.768 |
| 1a2p | 108 | 2.812 | 1.750 | 1.482 | 2.875 | 1.216 | 5.014 | 2.567 | 1.945 | 1.646 | 1.417 |
| 1hel | 129 | 2.453 | 1.767 | 1.635 | 2.856 | 0.708 | 2.387 | 1.159 | 1.135 | 1.029 | 1.529 |
| 1lz1 | 130 | 2.497 | 1.863 | 1.720 | 3.003 | 0.878 | 2.408 | 1.075 | 1.017 | 0.917 | 1.450 |
| anki | 156 | 2.239 | 2.450 | 2.628 | 1.390 | 1.486 | 2.535 | 1.642 | 1.486 | 1.452 | 0.696 |
| 1cus | 197 | 3.407 | 1.888 | 1.588 | 3.480 | 0.822 | 6.208 | 3.240 | 2.408 | 2.121 | 0.711 |
| 1inc | 240 | 6.118 | 3.513 | 2.770 | 5.053 | 1.605 | 6.929 | 2.924 | 1.964 | 2.058 | 0.721 |
| 1kvd | 280 | 2.050 | 1.683 | 1.834 | 1.775 | 0.815 | 4.127 | 1.923 | 1.357 | 1.456 | 0.803 |

Table 11: Average percentage error of the solvation energy values of 100 conformations for each protein ($\varepsilon_m = 1$).
[a]Beta3s is a three-stranded antiparallel $\beta$-sheet peptide [7, 8].

| | | PARAM19 | | | | | PARAM22 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FACTS | FACTS | FACTS | GBMV2 | GBgrid | FACTS | FACTS | FACTS | GBMV2 | GBgrid |
| PDB | residues | $\kappa = 4$ | $\kappa = 8$ | $\kappa = 12$ | $\kappa = 8$ | $\kappa = 8$ | $\kappa = 4$ | $\kappa = 8$ | $\kappa = 12$ | $\kappa = 8$ | $\kappa = 8$ |
| 1cb3 | 11 | 3.8 | 3.4 | 3.3 | 2.6 | 2.6 | 5.4 | 4.7 | 4.5 | 2.5 | 2.6 |
| bet1 | 14 | 4.8 | 4.2 | 4.1 | 3.5 | 3.6 | 5.1 | 4.3 | 4.2 | 3.7 | 3.9 |
| hlx1 | 17 | 5.2 | 4.8 | 4.8 | 4.1 | 4.4 | 6.2 | 5.7 | 5.6 | 4.5 | 4.7 |
| 1l2y | 20 | 6.2 | 5.5 | 5.4 | 4.4 | 3.9 | 7.2 | 6.1 | 5.8 | 4.0 | 3.8 |
| Beta3s[a] | 20 | 5.8 | 5.4 | 5.4 | 5.3 | 4.2 | 7.0 | 6.1 | 5.9 | 4.2 | 3.7 |
| 1f8a | 33 | 11.2 | 10.5 | 10.5 | 8.2 | 6.5 | 10.3 | 9.0 | 8.7 | 7.2 | 5.3 |
| 1abz | 38 | 14.4 | 13.6 | 13.4 | 8.3 | 10.0 | 11.2 | 10.6 | 10.7 | 8.9 | 8.9 |
| 1crn | 46 | 8.6 | 7.9 | 7.8 | 8.0 | 6.7 | 8.3 | 7.1 | 7.0 | 7.2 | 4.8 |
| ins2 | 51 | 12.6 | 12.0 | 11.9 | 9.1 | 8.3 | 10.2 | 9.8 | 10.1 | 8.0 | 6.3 |
| 1enh | 54 | 19.3 | 17.7 | 17.7 | 16.1 | 13.1 | 16.3 | 13.6 | 13.0 | 13.0 | 9.5 |
| 1pgb | 56 | 23.0 | 21.5 | 21.0 | 13.5 | 9.7 | 16.1 | 14.7 | 14.4 | 12.2 | 10.1 |
| 1shg | 57 | 19.4 | 18.6 | 18.7 | 16.0 | 12.0 | 12.5 | 12.3 | 12.6 | 9.4 | 8.9 |
| 1bpi | 58 | 19.2 | 18.4 | 18.3 | 15.3 | 9.6 | 15.2 | 14.4 | 14.4 | 12.4 | 6.9 |
| 1fmk | 59 | 19.3 | 17.8 | 17.3 | 14.5 | 7.7 | 13.8 | 12.1 | 11.8 | 10.4 | 9.0 |
| 2ptl | 61 | 20.1 | 18.4 | 18.0 | 14.3 | 12.5 | 14.0 | 13.2 | 13.4 | 10.6 | 9.3 |
| 2ci2 | 65 | 20.4 | 19.2 | 18.8 | 12.0 | 15.1 | 14.0 | 14.1 | 14.7 | 11.8 | 12.6 |
| 2a3d | 73 | 24.7 | 23.6 | 23.6 | 18.6 | 13.8 | 18.0 | 17.1 | 17.2 | 14.9 | 12.2 |
| 1ubq | 76 | 23.7 | 23.1 | 23.1 | 15.9 | 14.7 | 17.1 | 16.6 | 16.8 | 15.0 | 12.6 |
| 1pht | 83 | 29.9 | 28.4 | 28.0 | 23.5 | 14.7 | 17.8 | 17.4 | 17.9 | 14.5 | 11.5 |
| 1hdn | 85 | 18.2 | 17.5 | 17.7 | 16.3 | 14.5 | 17.1 | 15.4 | 15.2 | 13.5 | 13.6 |
| 1dvd | 98 | 28.1 | 26.5 | 26.4 | 22.8 | 18.9 | 20.0 | 19.0 | 19.3 | 16.4 | 13.7 |
| prph | 104 | 37.7 | 35.2 | 34.5 | 28.3 | 13.5 | 18.4 | 16.9 | 17.0 | 15.5 | 11.6 |
| 1a2p | 108 | 31.8 | 30.7 | 30.5 | 26.4 | 17.5 | 22.4 | 20.3 | 20.2 | 17.6 | 11.8 |
| 1hel | 129 | 40.0 | 38.3 | 38.1 | 39.2 | 21.6 | 32.5 | 30.5 | 30.0 | 27.6 | 14.7 |
| 1lz1 | 130 | 45.2 | 42.3 | 41.6 | 39.2 | 29.2 | 31.3 | 29.3 | 28.9 | 26.8 | 15.8 |
| anki | 156 | 94.9 | 91.1 | 89.6 | 59.1 | 37.3 | 34.9 | 30.7 | 29.6 | 31.3 | 21.8 |
| 1cus | 197 | 44.6 | 44.5 | 44.7 | 37.9 | 26.7 | 31.2 | 29.4 | 29.5 | 24.8 | 14.9 |
| 1inc | 240 | 45.4 | 44.6 | 44.7 | 40.3 | 28.1 | 37.1 | 34.8 | 34.7 | 29.5 | 17.0 |
| 1kvd | 280 | 86.9 | 83.7 | 83.1 | 58.6 | 38.1 | 34.2 | 30.1 | 29.4 | 27.0 | 19.8 |

Table 12: Average error in kcal/mol for the difference in electrostatic solvation energy ($\Delta\Delta G$) from *pairs* of protein conformations ($\varepsilon_m = 1$). For each protein, 4950 values of $\Delta\Delta G$ were calculated using 100 conformations. [a]Beta3s is a three-stranded antiparallel $\beta$-sheet peptide [7, 8].

| | FACTS $\kappa = 4$ | FACTS $\kappa = 8$ | FACTS $\kappa = 12$ | GBMV2 $\kappa = 8$ | GBgrid $\kappa = 8$ |
|---|---|---|---|---|---|
| **PARAM19** | | | | | |
| aver [%] | 2.121 | 1.433 | 1.360 | 2.371 | 2.036 |
| sig [%] | 1.644 | 1.260 | 1.229 | 1.343 | 1.315 |
| max [%] | 10.092 | 10.904 | 11.353 | 8.622 | 8.761 |
| **PARAM22** | | | | | |
| aver [%] | 3.538 | 1.896 | 1.560 | 1.285 | 2.287 |
| sig [%] | 1.958 | 1.274 | 1.142 | 1.030 | 1.480 |
| max [%] | 12.417 | 7.648 | 7.219 | 6.261 | 7.206 |

Table 13: Percentage error of electrostatic solvation energy for PARAM19 (top half) and PARAM22 (bottom half) from 2900 protein conformations (100 conformations from each of 29 trajectories), $\varepsilon_m = 1$

| | FACTS $\kappa = 4$ | FACTS $\kappa = 8$ | FACTS $\kappa = 12$ | GBMV2 $\kappa = 8$ | GBgrid $\kappa = 8$ |
|---|---|---|---|---|---|
| **PARAM19** | | | | | |
| aver [kcal/mol] | 26.359 | 25.117 | 24.897 | 20.045 | 14.431 |
| sig [kcal/mol] | 21.764 | 21.061 | 20.839 | 15.468 | 9.594 |
| max [kcal/mol] | 94.900 | 91.100 | 89.600 | 59.100 | 38.100 |
| **PARAM22** | | | | | |
| aver [kcal/mol] | 17.407 | 16.045 | 15.948 | 13.945 | 10.390 |
| sig [kcal/mol] | 9.536 | 8.824 | 8.720 | 8.310 | 4.955 |
| max [kcal/mol] | 37.100 | 34.800 | 34.700 | 31.300 | 21.800 |

Table 14: Error [kcal/mol] of solvation energy values from *pairs* of structures for PARAM19 (top half) and PARAM22 (bottom half) of 2900 conformations (100 conformations from each of 29 trajectories), $\varepsilon_m = 1$

| | | PARAM19 | | | | | PARAM22 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FACTS | FACTS | FACTS | GBMV2 | GBgrid | FACTS | FACTS | FACTS | GBMV2 | GBgrid |
| PDB | residue | $\kappa = 4$ | $\kappa = 8$ | $\kappa = 12$ | $\kappa = 8$ | $\kappa = 8$ | $\kappa = 4$ | $\kappa = 8$ | $\kappa = 12$ | $\kappa = 8$ | $\kappa = 8$ |
| 1cb3 | 11 | 0.995 | 0.995 | 0.995 | 0.997 | 0.998 | 0.989 | 0.991 | 0.992 | 0.998 | 0.999 |
| bet1 | 14 | 0.996 | 0.997 | 0.997 | 0.998 | 0.998 | 0.996 | 0.997 | 0.997 | 0.998 | 0.998 |
| hlx1 | 17 | 0.996 | 0.996 | 0.996 | 0.997 | 0.998 | 0.995 | 0.996 | 0.996 | 0.998 | 0.999 |
| 1l2y | 20 | 0.992 | 0.994 | 0.994 | 0.996 | 0.997 | 0.990 | 0.992 | 0.993 | 0.997 | 0.998 |
| Beta3s[a] | 20 | 0.992 | 0.993 | 0.993 | 0.992 | 0.996 | 0.988 | 0.990 | 0.991 | 0.996 | 0.998 |
| 1f8a | 33 | 0.994 | 0.995 | 0.995 | 0.997 | 0.999 | 0.995 | 0.996 | 0.996 | 0.998 | 0.999 |
| 1abz | 38 | 0.993 | 0.994 | 0.994 | 0.998 | 0.998 | 0.995 | 0.996 | 0.996 | 0.998 | 0.999 |
| 1crn | 46 | 0.994 | 0.995 | 0.995 | 0.995 | 0.997 | 0.995 | 0.996 | 0.996 | 0.996 | 0.998 |
| ins2 | 51 | 0.991 | 0.992 | 0.992 | 0.996 | 0.997 | 0.994 | 0.995 | 0.994 | 0.996 | 0.998 |
| 1enh | 54 | 0.993 | 0.994 | 0.994 | 0.995 | 0.997 | 0.994 | 0.996 | 0.996 | 0.997 | 0.999 |
| 1pgb | 56 | 0.991 | 0.992 | 0.993 | 0.997 | 0.999 | 0.996 | 0.997 | 0.997 | 0.998 | 0.999 |
| 1shg | 57 | 0.992 | 0.993 | 0.993 | 0.996 | 0.998 | 0.997 | 0.997 | 0.997 | 0.998 | 0.999 |
| 1bpi | 58 | 0.992 | 0.992 | 0.993 | 0.995 | 0.998 | 0.994 | 0.995 | 0.995 | 0.996 | 0.999 |
| 1fmk | 59 | 0.984 | 0.987 | 0.988 | 0.992 | 0.998 | 0.991 | 0.993 | 0.994 | 0.995 | 0.997 |
| 2ptl | 61 | 0.990 | 0.991 | 0.992 | 0.995 | 0.997 | 0.995 | 0.995 | 0.995 | 0.997 | 0.999 |
| 2ci2 | 65 | 0.991 | 0.992 | 0.992 | 0.998 | 0.998 | 0.997 | 0.997 | 0.996 | 0.998 | 0.999 |
| 2a3d | 73 | 0.994 | 0.995 | 0.995 | 0.997 | 0.999 | 0.997 | 0.997 | 0.997 | 0.998 | 0.999 |
| 1ubq | 76 | 0.992 | 0.993 | 0.993 | 0.996 | 0.998 | 0.996 | 0.996 | 0.996 | 0.997 | 0.999 |
| 1pht | 83 | 0.991 | 0.992 | 0.992 | 0.994 | 0.999 | 0.997 | 0.997 | 0.997 | 0.998 | 0.999 |
| 1hdn | 85 | 0.996 | 0.997 | 0.997 | 0.997 | 0.998 | 0.997 | 0.998 | 0.998 | 0.998 | 0.999 |
| 1dvd | 98 | 0.993 | 0.994 | 0.994 | 0.996 | 0.997 | 0.996 | 0.996 | 0.996 | 0.997 | 0.999 |
| prph | 104 | 0.972 | 0.975 | 0.976 | 0.985 | 0.997 | 0.994 | 0.995 | 0.994 | 0.995 | 0.998 |
| 1a2p | 108 | 0.985 | 0.987 | 0.987 | 0.991 | 0.997 | 0.992 | 0.994 | 0.994 | 0.995 | 0.998 |
| 1hel | 129 | 0.983 | 0.984 | 0.985 | 0.982 | 0.994 | 0.981 | 0.983 | 0.984 | 0.987 | 0.996 |
| 1lz1 | 130 | 0.985 | 0.987 | 0.988 | 0.989 | 0.994 | 0.990 | 0.991 | 0.992 | 0.993 | 0.998 |
| anki | 156 | 0.944 | 0.947 | 0.948 | 0.976 | 0.991 | 0.989 | 0.991 | 0.992 | 0.991 | 0.996 |
| 1cus | 197 | 0.987 | 0.987 | 0.986 | 0.990 | 0.995 | 0.992 | 0.993 | 0.993 | 0.995 | 0.998 |
| 1inc | 240 | 0.981 | 0.982 | 0.982 | 0.986 | 0.994 | 0.987 | 0.988 | 0.988 | 0.991 | 0.997 |
| 1kvd | 280 | 0.981 | 0.982 | 0.982 | 0.993 | 0.997 | 0.996 | 0.997 | 0.997 | 0.998 | 0.999 |

Table 15: Correlation coefficient of electrostatic solvation energy values of 100 conformations for each protein, $\varepsilon_m = 1$. The benchmark are the fdP data. [a]Beta3s is a three-stranded antiparallel $\beta$-sheet peptide [7, 8].

## 10 Energy in Solution

| | | PARAM19 | | | | | PARAM22 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FACTS | FACTS | FACTS | GBMV2 | GBgrid | FACTS | FACTS | FACTS | GBMV2 | GBgrid |
| PDB | residue | $\kappa = 4$ | $\kappa = 8$ | $\kappa = 12$ | $\kappa = 8$ | $\kappa = 8$ | $\kappa = 4$ | $\kappa = 8$ | $\kappa = 12$ | $\kappa = 8$ | $\kappa = 8$ |
| 1cb3 | 11 | 0.225 | 0.241 | 0.246 | 0.192 | 0.647 | 0.148 | 0.173 | 0.192 | 0.241 | 0.490 |
| bet1 | 14 | 0.112 | 0.141 | 0.173 | 0.086 | 0.547 | 0.166 | 0.079 | 0.080 | 0.123 | 0.337 |
| hlx1 | 17 | 0.103 | 0.131 | 0.143 | 0.162 | 0.590 | 0.106 | 0.092 | 0.100 | 0.243 | 0.496 |
| 1l2y | 20 | 0.167 | 0.099 | 0.100 | 0.286 | 0.388 | 0.192 | 0.092 | 0.080 | 0.100 | 0.310 |
| Beta3s[a] | 20 | 0.164 | 0.085 | 0.113 | 0.295 | 0.326 | 0.229 | 0.085 | 0.078 | 0.058 | 0.254 |
| 1f8a | 33 | 0.224 | 0.144 | 0.128 | 0.398 | 0.246 | 0.221 | 0.093 | 0.066 | 0.049 | 0.195 |
| 1abz | 38 | 0.142 | 0.117 | 0.113 | 0.231 | 0.362 | 0.148 | 0.085 | 0.072 | 0.057 | 0.231 |
| 1crn | 46 | 0.241 | 0.075 | 0.068 | 0.222 | 0.108 | 0.251 | 0.090 | 0.054 | 0.054 | 0.079 |
| ins2 | 51 | 0.217 | 0.103 | 0.086 | 0.156 | 0.169 | 0.258 | 0.141 | 0.110 | 0.049 | 0.103 |
| 1enh | 54 | 0.194 | 0.169 | 0.172 | 0.403 | 0.206 | 0.156 | 0.065 | 0.049 | 0.047 | 0.170 |
| 1pgb | 56 | 0.173 | 0.130 | 0.124 | 0.153 | 0.257 | 0.197 | 0.122 | 0.106 | 0.059 | 0.168 |
| 1shg | 57 | 0.165 | 0.145 | 0.141 | 0.273 | 0.222 | 0.195 | 0.126 | 0.109 | 0.043 | 0.183 |
| 1bpi | 58 | 0.210 | 0.168 | 0.161 | 0.369 | 0.137 | 0.172 | 0.089 | 0.070 | 0.053 | 0.135 |
| 1fmk | 59 | 0.217 | 0.109 | 0.097 | 0.156 | 0.162 | 0.247 | 0.127 | 0.096 | 0.083 | 0.073 |
| 2ptl | 61 | 0.165 | 0.109 | 0.103 | 0.195 | 0.209 | 0.210 | 0.118 | 0.096 | 0.051 | 0.162 |
| 2ci2 | 65 | 0.153 | 0.119 | 0.112 | 0.227 | 0.229 | 0.218 | 0.144 | 0.124 | 0.047 | 0.149 |
| 2a3d | 73 | 0.172 | 0.136 | 0.127 | 0.232 | 0.197 | 0.186 | 0.111 | 0.094 | 0.050 | 0.139 |
| 1ubq | 76 | 0.159 | 0.118 | 0.109 | 0.234 | 0.179 | 0.214 | 0.129 | 0.108 | 0.055 | 0.134 |
| 1pht | 83 | 0.182 | 0.130 | 0.119 | 0.151 | 0.185 | 0.221 | 0.150 | 0.136 | 0.081 | 0.078 |
| 1hdn | 85 | 0.167 | 0.090 | 0.079 | 0.200 | 0.139 | 0.236 | 0.136 | 0.109 | 0.049 | 0.121 |
| 1dvd | 98 | 0.162 | 0.118 | 0.108 | 0.189 | 0.139 | 0.211 | 0.138 | 0.118 | 0.054 | 0.115 |
| prph | 104 | 0.213 | 0.120 | 0.102 | 0.185 | 0.086 | 0.227 | 0.124 | 0.096 | 0.111 | 0.033 |
| 1a2p | 108 | 0.206 | 0.128 | 0.109 | 0.213 | 0.089 | 0.214 | 0.111 | 0.084 | 0.070 | 0.061 |
| 1hel | 129 | 0.208 | 0.150 | 0.139 | 0.243 | 0.059 | 0.122 | 0.059 | 0.058 | 0.052 | 0.079 |
| 1lz1 | 130 | 0.215 | 0.161 | 0.150 | 0.262 | 0.076 | 0.120 | 0.053 | 0.051 | 0.046 | 0.074 |
| anki | 156 | 0.230 | 0.251 | 0.269 | 0.144 | 0.155 | 0.160 | 0.103 | 0.094 | 0.093 | 0.045 |
| 1cus | 197 | 0.207 | 0.115 | 0.096 | 0.211 | 0.049 | 0.213 | 0.112 | 0.084 | 0.073 | 0.024 |
| 1inc | 240 | 0.277 | 0.159 | 0.126 | 0.230 | 0.072 | 0.200 | 0.085 | 0.057 | 0.060 | 0.021 |
| 1kvd | 280 | 0.136 | 0.110 | 0.119 | 0.119 | 0.053 | 0.159 | 0.075 | 0.053 | 0.056 | 0.031 |

Table 16: Average percentage error of the energy in solution (i.e., solvation energy plus vacuum energy) values of 100 conformations for each protein ($\varepsilon_m = 1$). [a]Beta3s is a three-stranded antiparallel $\beta$-sheet peptide [7, 8].

| | FACTS $\kappa = 4$ | FACTS $\kappa = 8$ | FACTS $\kappa = 12$ | GBMV2 $\kappa = 8$ | GBgrid $\kappa = 8$ |
|---|---|---|---|---|---|
| **PARAM19** | | | | | |
| aver [%] | 0.186 | 0.133 | 0.129 | 0.221 | 0.217 |
| sig [%] | 0.121 | 0.114 | 0.116 | 0.119 | 0.175 |
| max [%] | 1.158 | 1.130 | 1.108 | 0.741 | 0.874 |
| **PARAM22** | | | | | |
| aver [%] | 0.193 | 0.107 | 0.091 | 0.076 | 0.155 |
| sig [%] | 0.087 | 0.070 | 0.068 | 0.069 | 0.132 |
| max [%] | 0.608 | 0.488 | 0.510 | 0.467 | 0.678 |

Table 17: Percentage error of energy in solution values for PARAM19 (top half) and PARAM22 (bottom half) of 2900 conformations (100 conformations from each of 29 trajectories), $\varepsilon_m = 1$

| | | PARAM19 | | | | | PARAM22 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FACTS | FACTS | FACTS | GBMV2 | GBgrid | FACTS | FACTS | FACTS | GBMV2 | GBgrid |
| PDB | residue | $\kappa = 4$ | $\kappa = 8$ | $\kappa = 12$ | $\kappa = 8$ | $\kappa = 8$ | $\kappa = 4$ | $\kappa = 8$ | $\kappa = 12$ | $\kappa = 8$ | $\kappa = 8$ |
| 1cb3 | 11 | 0.976 | 0.981 | 0.982 | 0.988 | 0.992 | 0.825 | 0.860 | 0.872 | 0.970 | 0.975 |
| bet1 | 14 | 0.984 | 0.988 | 0.989 | 0.991 | 0.995 | 0.940 | 0.960 | 0.962 | 0.975 | 0.972 |
| hlx1 | 17 | 0.983 | 0.986 | 0.986 | 0.988 | 0.994 | 0.925 | 0.937 | 0.940 | 0.976 | 0.980 |
| 1l2y | 20 | 0.960 | 0.970 | 0.972 | 0.979 | 0.988 | 0.866 | 0.901 | 0.911 | 0.963 | 0.976 |
| Beta3s[a] | 20 | 0.973 | 0.977 | 0.977 | 0.972 | 0.987 | 0.884 | 0.911 | 0.917 | 0.963 | 0.978 |
| 1f8a | 33 | 0.976 | 0.981 | 0.982 | 0.988 | 0.994 | 0.941 | 0.955 | 0.958 | 0.971 | 0.986 |
| 1abz | 38 | 0.979 | 0.982 | 0.982 | 0.994 | 0.994 | 0.963 | 0.969 | 0.969 | 0.981 | 0.987 |
| 1crn | 46 | 0.980 | 0.983 | 0.983 | 0.982 | 0.989 | 0.948 | 0.961 | 0.963 | 0.963 | 0.985 |
| ins2 | 51 | 0.968 | 0.970 | 0.970 | 0.983 | 0.990 | 0.950 | 0.953 | 0.951 | 0.973 | 0.986 |
| 1enh | 54 | 0.986 | 0.987 | 0.987 | 0.992 | 0.995 | 0.972 | 0.979 | 0.980 | 0.981 | 0.988 |
| 1pgb | 56 | 0.971 | 0.975 | 0.976 | 0.993 | 0.997 | 0.957 | 0.965 | 0.966 | 0.979 | 0.991 |
| 1shg | 57 | 0.970 | 0.973 | 0.973 | 0.984 | 0.992 | 0.967 | 0.967 | 0.965 | 0.981 | 0.986 |
| 1bpi | 58 | 0.973 | 0.976 | 0.976 | 0.986 | 0.995 | 0.948 | 0.954 | 0.954 | 0.967 | 0.990 |
| 1fmk | 59 | 0.976 | 0.979 | 0.980 | 0.988 | 0.996 | 0.956 | 0.967 | 0.970 | 0.979 | 0.989 |
| 2ptl | 61 | 0.963 | 0.968 | 0.970 | 0.984 | 0.990 | 0.957 | 0.964 | 0.964 | 0.979 | 0.990 |
| 2ci2 | 65 | 0.973 | 0.975 | 0.976 | 0.994 | 0.994 | 0.975 | 0.978 | 0.977 | 0.986 | 0.988 |
| 2a3d | 73 | 0.976 | 0.978 | 0.979 | 0.987 | 0.995 | 0.958 | 0.961 | 0.960 | 0.970 | 0.984 |
| 1ubq | 76 | 0.977 | 0.978 | 0.978 | 0.991 | 0.993 | 0.959 | 0.962 | 0.961 | 0.973 | 0.986 |
| 1pht | 83 | 0.964 | 0.969 | 0.971 | 0.980 | 0.992 | 0.955 | 0.959 | 0.957 | 0.970 | 0.984 |
| 1hdn | 85 | 0.984 | 0.985 | 0.985 | 0.989 | 0.993 | 0.962 | 0.970 | 0.971 | 0.980 | 0.988 |
| 1dvd | 98 | 0.979 | 0.982 | 0.982 | 0.989 | 0.993 | 0.960 | 0.964 | 0.964 | 0.976 | 0.989 |
| prph | 104 | 0.963 | 0.968 | 0.969 | 0.981 | 0.995 | 0.971 | 0.975 | 0.975 | 0.980 | 0.990 |
| 1a2p | 108 | 0.978 | 0.979 | 0.980 | 0.987 | 0.995 | 0.965 | 0.971 | 0.971 | 0.976 | 0.990 |
| 1hel | 129 | 0.966 | 0.968 | 0.968 | 0.967 | 0.988 | 0.926 | 0.934 | 0.935 | 0.946 | 0.984 |
| 1lz1 | 130 | 0.958 | 0.963 | 0.965 | 0.974 | 0.982 | 0.927 | 0.937 | 0.939 | 0.946 | 0.982 |
| anki | 156 | 0.919 | 0.927 | 0.931 | 0.969 | 0.985 | 0.956 | 0.963 | 0.965 | 0.959 | 0.979 |
| 1cus | 197 | 0.972 | 0.971 | 0.970 | 0.980 | 0.990 | 0.948 | 0.953 | 0.953 | 0.966 | 0.988 |
| 1inc | 240 | 0.932 | 0.934 | 0.934 | 0.952 | 0.977 | 0.892 | 0.905 | 0.906 | 0.934 | 0.977 |
| 1kvd | 280 | 0.947 | 0.951 | 0.951 | 0.981 | 0.988 | 0.960 | 0.970 | 0.971 | 0.977 | 0.987 |

Table 18: Correlation coefficient of electrostatic energy in solution (i.e., solvation energy plus vacuum energy) values of 100 conformations for each protein, $\varepsilon_m = 1$. The benchmark are the fdP data. [a]Beta3s is a three-stranded antiparallel $\beta$-sheet peptide [7, 8].

# Chapter 6

# Fast Protein Folding on Downhill Energy Landscape

**FOR THE RECORD**

# Fast protein folding on downhill energy landscape

ANDREA CAVALLI, URS HABERTHÜR, EMANUELE PACI, AND AMEDEO CAFLISCH
Department of Biochemistry, University of Zürich, CH-8057 Zürich, Switzerland

(RECEIVED March 4, 2003; FINAL REVISION May 1, 2003; ACCEPTED May 1, 2003)

**Abstract**

Proteins fold in a time range of microseconds to minutes despite the large amount of possible conformers. Molecular dynamics simulations of a three-stranded antiparallel β-sheet peptide (for a total of 12.6 μsec and 72 folding events) show that at the melting temperature the unfolded state ensemble contains many more conformers than those sampled during a folding event.

**Keywords:** Protein folding; Levinthal's paradox; molecular dynamics; unfolded state; β-sheet; folding rate

Proteins are complex molecules with many degrees of freedom. Their ability to fold to a unique three-dimensional structure in a time range of microseconds to minutes seems to be at odds with the large amount of possible conformers (Dill and Chan 1997; Karplus 1997). One argument against this apparent conundrum (Levinthal's paradox) is based on the results of a "toy" model of proteins, a string of 27 beads positioned at sites of a cubic lattice where beads interact only if they are nonbonded nearest neighbors and long-range interactions are neglected. Because the number of accessible conformations is of the order of $10^{16}$, folding of the lattice model in about $10^7$ Monte Carlo moves has suggested that it is possible to reach the folded state after searching through only a minute fraction of the denaturated state ensemble (Leopold et al. 1992; Dinner et al. 2000; Dinner and Karplus 2001). On the basis of explicit solvent molecular dynamics simulations of structured peptides (lasting up to 200 nsec) it has been proposed recently that the number of conformers that characterize the denaturated state is only on the order of $10^9$ for a 100-residue protein that folds on a millisecond time scale (van Gunsteren et al. 2001a). Furthermore, the significance of the results obtained by lattice simulations that model only short-range interactions (Dinner and Karplus 2001) has been questioned (van Gunsteren et al. 2001b).

Reprint requests to: Amedeo Caflisch, Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland; e-mail: caflisch@bioc.unizh.ch; fax: (41-1) 635-68-62.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.0366103.

In this paper we show that the denatured state ensemble of a small protein cannot be characterized by a small number of statistically relevant conformations. Folding occurs through the exploration of a small number of conformations, and different conformations are sampled in different folding events. Beta3s is a designed 20-residue sequence whose solution conformation has been investigated by NMR spectroscopy (de Alba et al. 1999). The NMR data indicate that beta3s in aqueous solution forms a monomeric (up to 1 mM concentration) triple-stranded antiparallel β-sheet (Fig. 1, inset), in equilibrium with the random coil (de Alba et al. 1999). We have shown previously that in implicit solvent (Ferrara et al. 2002) molecular dynamics simulations beta3s folds reversibly to the NMR solution conformation, irrespective of the starting conformation (Ferrara and Caflisch 2000; Cavalli et al. 2002). Recently, four additional molecular dynamics simulations of beta3s were performed at 330 K for a total simulation time of 12.6 μsec. The length of each simulation (2.7 μsec, 2.7 μsec, 2.8 μsec, and 4.4 μsec) is more than 30 times longer than the average folding or unfolding time (about 85 nsec each), which are similar because at 330 K the folded and unfolded states are equally populated. At 330 K the peptide is within 2.5 Å $C_\alpha$ root mean square deviation (RMSD) from the folded conformation about 48% of the time. Figure 1 shows the results of a cluster analysis based on $C_\alpha$ RMSD. There are more than 15,000 conformers (cluster centers) and it is evident that a plateau has not been reached within the 12.6 μsec of simulation time. However, the number of significantly populated clusters (see Ferrara and Caflisch 2001 for a detailed description) converges already within 2 μsec.

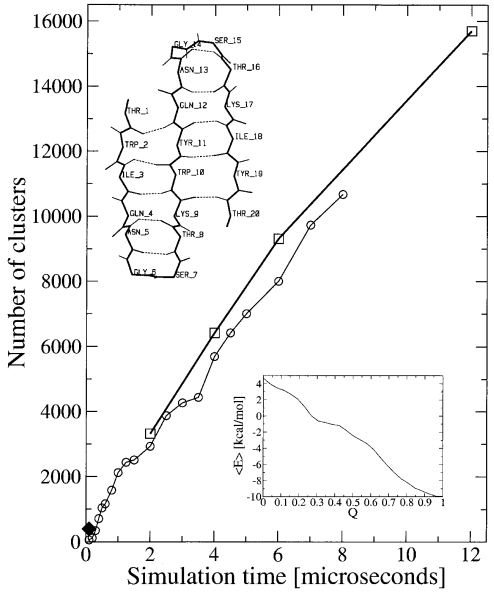


**Figure 1.** Number of clusters as a function of time. The "leader" clustering procedure was used with a total of 120,000 snapshots saved every 0.1 nsec (thick line, squares). The clustering algorithm, which uses the $C_\alpha$ RMSD values between all pairs of structures, was used only for the first 8 μsec (80,000 snapshots) because of the computational requirements (thin line, circles). (Diamond) Average number of conformers sampled during the folding time, which is defined as the average time interval between successive unfolding and refolding events. (*Inset, top*) A backbone representation of the folded state of beta3s with main chain hydrogen bonds as broken lines; (*inset, bottom*) average effective energy as a function of the fraction of native contacts Q, which are defined in Ferrara and Caflisch 2000.

Hence, the simulation-length dependence of the total number of clusters is dominated by the small ones. At each simulation interval between an unfolding event and the successive refolding event additional conformations are sampled (Fig. 2). More than 90% of the unfolded state conformations are in small clusters (each containing < 0.1% of the saved snapshots) and the total number of small clusters does not reach a plateau within 12.6 μsec. Note that there is also a monotonic growth with simulation time of the number of snapshots in the folded-state cluster. After 12.6 μsec (and also within each of the four trajectories) the system has sampled at the equilibrium of folded and unfolded states despite the fact that a large part of the denaturated state ensemble has not yet been explored. In fact, the average folding time converges to a value around 85 nsec, which shows that the length of each simulation is much larger than the relaxation time of the slowest conformational change. Interestingly, in the average folding time of about 85 nsec beta3s visits < 400 clusters (diamond in Fig. 1). This is only a small fraction of the total number of conformers in the denaturated state. It is possible to reconcile the fast folding with the large conformational space by analyzing the effective energy, which includes all of the contributions to the free energy except for the configurational entropy of the protein (Dinner et al. 2000; Ferrara and Caflisch 2000). Fast folding of beta3s is consistent with the monotonically decreasing profile of the effective energy (Fig. 1, inset). Despite the large number of conformers in the denaturated state ensemble, the protein chain efficiently finds its way to the folded state because native-like interactions are on average more stable than non-native ones.

In conclusion, we have shown using an atomic model of a small protein that the unfolded state ensemble at the melting temperature is a large collection of conformers differing among each other, in agreement with previous high-temperature molecular dynamics simulations (Wong et al. 2000; Shea and Brooks 2001). The energy "bias" that makes fast folding possible does not imply that the unfolded state ensemble is made up of a small number of statistically relevant conformations. The simulation results provide further evidence that the number of denaturated state conformations is orders of magnitudes larger than the conformers sampled during a folding event.

## Materials and methods

The molecular dynamics simulations and part of the analysis of the trajectories were performed with CHARMM (Brooks et al. 1983). Beta3s was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field; Brooks et al. 1983). An implicit model based on the solvent-accessible surface was used to describe the main effects of the aqueous solvent on the solute (Ferrara et al. 2002). The CHARMM PARAM19 default cutoffs for long-range interactions were used, that is, a shift function (Brooks et al. 1983)

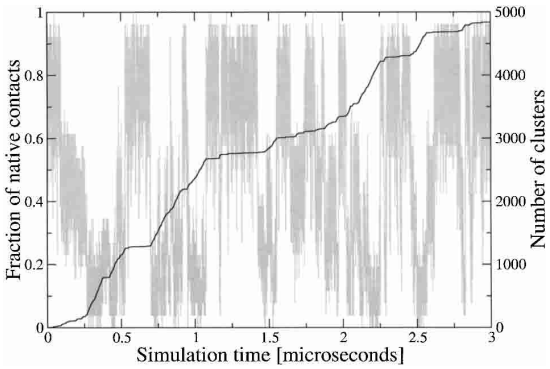


**Figure 2.** Time series of the fraction of native contacts Q (gray line, axis labels on the *left*) and total number of clusters (thick line, axis labels on the *right*) along one of the four trajectories. The plot shows that the number of clusters grows in the simulation intervals during which beta3s is in the unfolded state, i.e., Q values close to zero.

was employed with a cutoff at 7.5 Å for both the electrostatic and van der Waals terms. This cutoff length was chosen to be consistent with the parameterization of the force-field and implicit solvation model. The model is not biased toward any particular secondary structure type. In fact, exactly the same force field and implicit solvent model have been used recently in molecular dynamics simulations of folding of structured peptides (α-helices and β-sheets) ranging in size from 15 to 31 residues (Ferrara and Caflisch 2000, 2001; Hiltpold et al. 2000), and small proteins of about 60 residues (Gsponer and Caflisch 2001, 2002). Despite the lack of friction attributable to the absence of explicit water molecules, the implicit solvent model yields a separation of time scales consistent with experimental data near room temperature: Helices fold in about 1 nsec (Ferrara et al. 2000; ≈0.1 μsec, experimentally [Eaton et al., 2000]), β-hairpins in about 10 nsec (Ferrara et al. 2000; ≈1μsec [Eaton et al. 2000]), and triple-stranded β-sheets in about 100 nsec (≈10μsec experimentally; de Alba et al. 1999).

The trajectories were started from the folded state with different initial assignment of the velocities. The temperature was kept constant at 330 K by weak coupling to an external bath with a coupling constant of 5 psec. The value of 330 K is close to the melting temperature in the model (Cavalli et al. 2002). The SHAKE algorithm (Ryckaert et al. 1977) was used to fix the length of the covalent bonds involving hydrogen atoms, which allows an integration time step of 2 fsec.

The fraction of native contacts Q is a progress variable whose time dependence is used to monitor folding/unfolding events (Ferrara and Caflisch 2000). A folding event is considered completed when Q reaches a value larger than 0.85 ($Q > 22/26$), while an unfolding event is considered completed when Q drops below 0.15 ($Q < 4/26$; Ferrara and Caflisch 2000). The folding time is defined as the temporal interval between the first time point with $Q > 22/26$ and the first time point with $Q < 4/26$. The unfolding time is defined analogously, that is, the interval between the first time point with $Q < 4/26$ and the first time point with $Q > 22/26$.

The method for cluster analysis ("leader" algorithm) is based on structural similarity. The first conformation along a trajectory is defined as the center of the first cluster. The remaining conformations are iteratively added to the cluster whose center has the lowest $C_\alpha$ RMSD if the $C_\alpha$ RMSD is smaller than a cutoff of 2 Å. If the closest cluster center deviates more, the conformation becomes the center of a new cluster. To estimate the statistical error the clustering was repeated several times. For this purpose, the four simulations were concatenated and the resulting composite trajectory was divided in subintervals of equal length (e.g., the number of clusters sampled in 4 μsec is calculated three times on the intervals 0–4 μsec, 4–8 μsec, and 8–12 μsec). The statistical error in the number of clusters is about twice the size of the square symbols in Figure 1. To show that the overall behavior does not depend on the clustering procedure a different clustering algorithm was also used. It evaluates the $C_\alpha$ RMSD for *each* pair of structures (Daura et al. 1999). Both clustering procedures gave a similar simulation-length dependence of the number of clusters for $C_\alpha$ RMSD cutoff values ranging from 1.5 to 2.5 Å.

## Acknowledgments

We thank J. Gsponer for interesting discussions and comments to the manuscript. We also thank A. Widmer (Novartis Pharma, Basel, Switzerland) for providing the molecular modeling program Wit!P, which was used for visual and cluster analysis of the trajectories. We are grateful to M. Schaefer (Syngenta, Basel, Switzerland) for providing the program used for the clustering with RMSD between all pairs of structures. This work was supported by the Swiss National Competence Center in Structural Biology (NCCR) and the Swiss National Science Foundation (grant no. 31–64968.01 to A.C.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

## References

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization and dynamics calculations. *J. Comp. Chem.* **4:** 187–217.

Cavalli, A., Ferrara, P., and Caflisch, A. 2002. Weak temperature dependence of the free energy surface and folding pathways of structured peptides. *Proteins* **47:** 305–314.

Daura, X., van Gunsteren, W.F., and Mark, A.E. 1999. Folding-unfolding thermodynamics of a β-heptapeptide from equilibrium simulations. *Proteins* **34:** 269–280.

de Alba, E., Santorio, J., Rico, M., and Jimenez, M.A. 1999. De novo design of a monomeric three-stranded antiparallel β-sheet. *Protein Sci.* **8:** 854–865.

Dill, K.A. and Chan, H.S. 1997. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4:** 10–19.

Dinner, A.R. and Karplus, M. 2001. Comment on the communication "The key to solving the protein-folding problem lies in an accurate description of the denatured state" by van Gunsteren et al. *Angew. Chem. Int. Ed.* **40:** 4615–4616.

Dinner, A.R., Sali, A., Smith, L.J., Dobson, C.M., and Karplus, M. 2000. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* **25:** 331–339.

Eaton, W.A., Muñoz, V., Hagen, S.J., Jas, G.S., Lapidus, L.J., Henry, E.R., and Hofrichter, J. 2000. Fast kinetics and mechanisms in protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **29:** 327–359.

Ferrara, P. and Caflisch, A. 2000. Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc. Natl. Acad. Sci.* **97:** 10780–10785.

———. 2001. Native topology or specific interactions: What is more important for protein folding? *J. Mol. Biol.* **306:** 837–850.

Ferrara, P., Apostolakis, J., and Caflisch, A. 2000. Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. *J. Phys. Chem. B* **104:** 5000–5010.

———. 2002. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins* **46:** 24–33.

Gsponer, J., and Caflisch, A. 2001. Role of native topology investigated by multiple unfolding simulations of four SH3 domains. *J. Mol. Biol.* **309:** 285–298.

———. 2002. Molecular dynamics simulations of protein folding from the transition state. *Proc. Natl. Acad. Sci.* **99:** 6719–6724.

Hiltpold, A., Ferrara, P., Gsponer, J., and Caflisch, A. 2000. Free energy surface of the helical peptide Y(MEARA)(6). *J. Phys. Chem. B* **104:** 10080–10086.

Karplus, M. 1997. The Levinthal paradox: Yesterday and today. *Fold Des.* **2:** S69-S75.

Leopold, P.E., Montal, M., and Onuchic, J.N. 1992. Protein folding funnels: A kinetic approach to the sequence-structure. *Proc. Natl. Acad. Sci.* **89:** 8721–8725.

Ryckaert, J.P., Ciccotti, G., and Berendsen, H.J.C. 1977. Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **23:** 327–341.

Shea, J.E. and Brooks, C.L. 2001. From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem* **52:** 499–535.

van Gunsteren, W.F., Bürgi, R., Peter, C., and Daura, X. 2001a. The key to solving the protein-folding problem lies in an accurate description of the denatured state. *Angew. Chem. Int. Ed.* **40:** 351–355.

———. 2001b. Reply. *Angew. Chem. Int. Ed.* **40:** 4616–4618.

Wong, K.B., Clarke, J., Bond, C.J., Neira, J.L., Freund, S.M., Fersht, A.R., and Daggett, V. 2000. Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding. *J. Mol. Biol.* **296:** 1257–1282.

# Chapter 7

# The Role of Side Chain Interactions in the Early Steps of Aggregation: Molecular Dynamics Simulations of an Amyloid-Forming Peptide from the Yeast Prion Sup35

# The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35

**Jörg Gsponer, Urs Haberthür, and Amedeo Caflisch***

Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

Edited by R. Stephen Berry, University of Chicago, Chicago, IL, and approved January 28, 2003 (received for review September 2, 2002)

**Understanding the early steps of aggregation at atomic detail might be crucial for the rational design of therapeutics preventing diseases associated with amyloid deposits. In this paper, aggregation of the heptapeptide GNNQQNY, from the N-terminal prion-determining domain of the yeast protein Sup35, was studied by 20 molecular dynamics runs for a total simulation time of 20 $\mu$s. The simulations generate in-register parallel packing of GNNQQNY $\beta$-strands that is consistent with x-ray diffraction and Fourier transform infrared data. The statistically preferred aggregation pathway does not correspond to a purely downhill profile of the energy surface because of the presence of enthalpic barriers that originate from out-of-register interactions. The parallel $\beta$-sheet arrangement is favored over the antiparallel because of side-chain contacts; in particular, stacking interactions of the tyrosine rings and hydrogen bonds between amide groups. No ordered aggregation was found in control simulations with the mutant sequence SQNGNQQRG in accord with experimental data and the strong sequence dependence of aggregation.**

protein aggregation | misfolding | energy landscape

Amyloid fibrils are highly ordered protein aggregates associated with severe human disorders including Alzheimer's disease, type II diabetes, systemic amyloidosis, and transmissible spongiform encephalopathies (1, 2). The soluble precursors of the amyloidogenic proteins do not share any sequence homology or common fold. However, x-ray diffraction data indicate a cross-$\beta$ structure for all amyloid fibrils (3, 4). These findings suggest that key steps in the aggregation process may be common to all amyloidogenic proteins. Despite the medical relevance of amyloidosis, many important questions about the formation of ordered aggregates remain unanswered. What energetic contributions stabilize the species formed early in the aggregation process? In particular, what is the role of side-chain interactions and what are the most favorable side-chain arrangements? How sensitive is amyloid formation to small changes in the amino acid sequence?

There have been several lattice studies on aggregation in proteins. These simplified models have allowed for the investigation of the relevance of aggregation on the folding process (5) and how interaction potentials affect the properties of aggregation-prone proteins (6). Harrison *et al.* (7) have shown that less stable proteins have a greater chance of assuming alternative native states as multimers. Molecular dynamics (MD) simulations of aggregation have been performed by using a three-bead backbone and a single-bead side-chain model (8). Although this simplified model has allowed the simulation of the competition between folding and aggregation for two four-helix bundles, it is probably not possible to extract detailed information on energetics and sequence dependence. Recently, MD simulations of atomic models of amyloidogenic peptides have been performed with an implicit treatment of the solvent (9) and explicit water molecules (10, 11). In the former, the role of complex environments on the stabilization of intermolecular hydrogen bonds was investigated (9). The simulations of oligomers of Alzheimer's amyloid peptides in explicit water indicate that $A\beta_{16-22}$ aggregates with an antiparallel $\beta$-sheet orientation in agreement with solid state NMR data and that $A\beta_{16-35}$ cannot form linear parallel $\beta$-sheets because of unfavorable polar contacts (11).

The heptapeptide GNNQQNY from the yeast prion Sup35 (residues 7–13) displays the same amyloid properties as full-length Sup35, including cooperative kinetics of aggregation, fibril formation, binding of the dye Congo red, and the cross-$\beta$ x-ray diffraction pattern (12). The experimental evidence on GNNQQNY indicates that the amyloid-forming nucleus of a protein might consist of only a short segment of the entire chain. Furthermore, it has recently been shown that cytotoxicity is more pronounced for the early aggregates than for highly organized fibrillar structures (13). In this report, the free-energy surface of the very early steps of aggregation and the role of cross-strand side-chain interactions are investigated by implicit solvent (14) MD simulations of a trimer of the heptapeptide GNNQQNY. A set of mutant peptides are also simulated to explore the sensitivity to amino acid sequence.

## Materials and Methods

**Model.** The MD simulations and part of the analysis of the trajectories were performed with the CHARMM program (15). The peptide was modeled by the CHARMM PARAM19 force field, i.e., by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (15). An implicit model based on the solvent-accessible surface was used to describe the main effects of the aqueous solvent on the solute (14). The CHARMM PARAM19 default cutoffs for long-range interactions were used, i.e., a shift function (15) was used with a cutoff at 7.5 Å for both the electrostatic and van der Waals terms. This cutoff length was chosen to be consistent with the parameterization of the force field and implicit solvation model. The model is not biased toward any particular secondary structure type. In fact, exactly the same force field and implicit solvent model have been used recently in MD simulations of folding of structured peptides ($\alpha$-helices and $\beta$-sheets) ranging in size from 15 to 31 residues (16, 17, 18) and small proteins of about 60 residues (19, 20). Despite the lack of friction due to the absence of explicit water molecules, the implicit solvent model yields a separation of time scales consistent with experimental data: helices fold in $\approx$1 ns (21) [$\approx$100 ns experimentally (22)], $\beta$-hairpins fold in $\approx$10 ns (21) [$\approx$1 $\mu$s (22)], and triple-stranded $\beta$-sheets fold in $\approx$100 ns (23) [$\approx$10 $\mu$s (24)].

**Simulations.** All simulations were performed with three replicas starting from random conformations, positions, and orientations. In the initial random positions there was no intermolecular contact, i.e., the peptides were separated in space. The temperature was kept close to 330 K by weak coupling to an external

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: MD, molecular dynamics.

*To whom correspondence should be addressed. E-mail: caflisch@bioc.unizh.ch.



www.pnas.org/cgi/doi/10.1073/pnas.0835307100

**Table 1. Simulations performed**

| Peptide sequence | No. of simulations | Length, $\mu$s | No. of IP3* aggregation events | No. of IA3† aggregation events |
|---|---|---|---|---|
| GNNQQNY | 20 | 1 | 25 (14.9)‡ | 7 (3.1) |
| GNNQQNA | 3 | 2 | 5 (2.9) | 5 (3.1) |
| GNNQQNG | 3 | 2 | 6 (3.6) | 7 (2.8) |
| GNNQQN | 2 | 2 | 2 (1.3) | 3 (1.0) |
| GNNQQNNG | 3 | 1 | 1 (6.5) | 4 (11.0) |
| SQNGNQQRG | 3 | 2 | 0 | 0 |
| SENGNQQRG | 3 | 1 | 0 | 0 |

Each trajectory simulates three replicas of a given sequence.
*Three-stranded parallel in-register aggregates.
†Three-stranded antiparallel in-register aggregates.
‡The average time (ns) the replicas remained aggregated in IP3 and IA3 is given in parentheses.

bath with a coupling constant of 5 ps. A temperature of 330 K was chosen to get a statistically significant number of aggregation and disaggregation events in the time scale of the simulations. The MMFP option (25) of CHARMM was used to prevent the peptides from leaving a sphere of 150-Å diameter. The SHAKE algorithm (26) was used to fix the length of the covalent bonds involving hydrogen atoms, allowing an integration time step of 2 fs. The nonbonded interactions were updated every 10 dynamics steps and coordinate frames were saved every 20 ps for a total of 50,000 conformations per $\mu$s. A 1-$\mu$s run requires $\approx$20 days on a 1.4-GHz Athlon processor.

**Progress Variables.** The conformations sampled at 330 K were used to define the aggregation contacts between in-register and out-of-register strands. Backbone and side-chain contacts were considered to be present if the $C_\alpha$ atoms were within 5.5 Å and the center of mass of the side chains was within 6.0 Å.

**Normalized Frequency.** The normalized frequency of forming IP2 aggregates is given by

$$\frac{N_{\text{IP2-DA}}}{t_{\text{DA}}}, \qquad [1]$$

where $N_{\text{IP2-DA}}$ is the number of transitions between the disordered aggregates DA and the double stranded in-register aggregate IP2, and $t_{\text{DA}}$ is the time during which the three peptides do not form ordered aggregates, i.e., $Q_a \leq 0.2$ and $Q_p \leq 0.2$.

## Results and Discussion

**Strategy to Simulate Aggregation.** Because the major goal of this report is to study the early steps of aggregation, MD simulations were performed with three peptide replicas. Although it is not known experimentally whether three peptides form a stable "nucleus," the small number of replicas kept the complexity of the system and the CPU requirements low. Simulations were started from random conformations, positions, and orientations of the three replicas and carried out for 1 or 2 $\mu$s at 330 K (Table 1).

**Aggregation Events.** In 20 simulations, the three replicas of the GNNQQNY heptapeptides formed 25 times an in-register parallel $\beta$-sheet (IP3), irrespective of the starting conformation of the peptides and their relative position and orientation (Fig. 1). The observed parallel packing of the $\beta$-strands is consistent with x-ray diffraction and Fourier transform infrared (FTIR) data (12). Yet, it is important to note that the experimental data supporting a parallel arrangement are not conclusive and in particular FTIR can be misleading on this point. Furthermore,

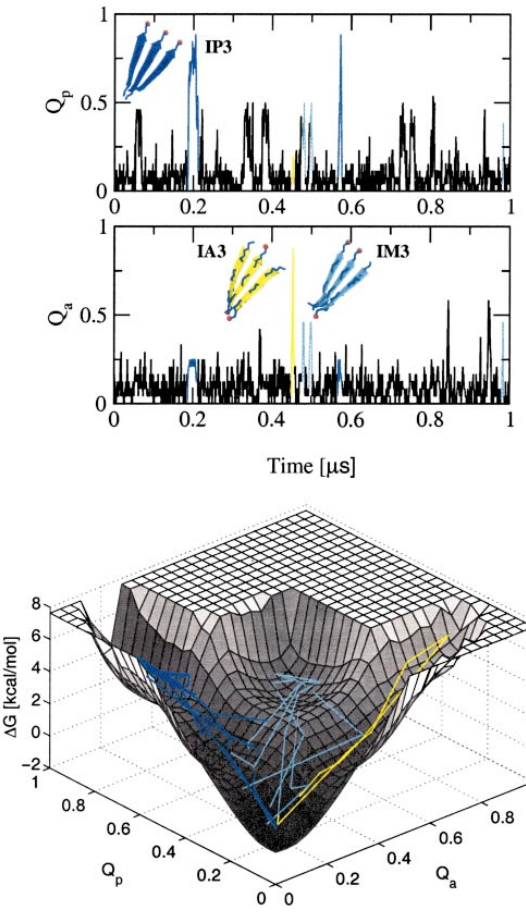


**Fig. 1.** (*Upper*) Time dependence of the fraction of in-register parallel contacts $Q_p$ and in-register antiparallel contacts $Q_a$ for one trajectory of the GNNQQNY peptide. Aggregation events to IP3, IM3, and IA3 are shown in blue, cyan, and yellow, respectively. (*Lower*) Projection of the aggregation events onto the free-energy surface (for the construction of the free-energy surface, see the text and the caption of Fig. 2).

the structure in the microcrystals is not necessarily the same as that in the amyloid fibrils. The average spacing between the $\beta$-strands in IP3 is 4.90 $\pm$ 0.12 Å (4.55 $\pm$ 0.14 Å after minimization). The in-register parallel $\beta$-sheet organization juxtaposes the polar residues asparagine and glutamine of neighboring peptide chains, as well as the aromatic rings of the tyrosines. This configuration enables the formation of, on average, 10 hydrogen bonds. The IP3 conformations sampled by MD are consistent with the suggestion of Balbirnie *et al.* (12) that a large number of side-chain hydrogen bonds contribute to the high density and stability observed for microcrystals of the GNNQQNY heptapeptides.

The three replicas also formed seven times an in-register antiparallel (IA3) and 42 times an in-register mixed parallel–antiparallel (IM3) $\beta$-sheet during the 20 $\mu$s of simulation time (Fig. 1). These two types of in-register aggregates have a reduced kinetic stability compared with IP3. The latter is stable for an average of 14.9 ns, whereas IM3 and IA3 disaggregate, on average, after 8.0 and 3.1 ns, respectively. The antiparallel

BIOPHYSICS

alignment of the strands forbids, partially in IM3 and completely in IA3, the interactions between the aromatic rings of the tyrosines. Moreover, the average number of side-chain hydrogen bonds is reduced to seven and five in IM3 and IA3, respectively. These findings are in agreement with theoretical results indicating that cross-strand interactions between side chains are required for the formation of stable $\beta$-sheets (27). The average distance between the antiparallel strands is $4.85 \pm 0.15$ Å ($4.55 \pm 0.38$ Å after minimization), which is similar to the average observed for the parallel aggregate.

A total of 257 partial aggregation events to an in-register parallel (IP2) and 142 to an in-register antiparallel double-stranded $\beta$-sheet (IA2) occurred. The third strand did either not interact with the two-stranded aggregate or was forming out-of-register interactions with it. Hence, not only in-register but also out-of-register aggregates were observed in the simulations. A mixed $\beta$-sheet consisting of two parallel in-register strands and a third antiparallel one that is displaced by one residue (OM3) constitutes a special subgroup of IP2. OM3 is the only long-lived out-of-register aggregate. Thirty-nine aggregation events to OM3 were observed. With an average disaggregation time of 9.3 ns, its kinetic stability is even slightly higher than that of IM3. All other types of out-of-register aggregates with a displacement of more than one residue in any chain were short-lived and dissolved quickly. Moreover, a cluster analysis showed that IP3, IM3, and OM3 are the only highly populated three-stranded aggregates.

Several runs with control peptides (for a total of 28 $\mu$s) were performed to test the reliability of the simulation protocol (Table 1). Experimental studies on the nonapeptide, SQNGNQQRG (Sup35 residues 17–25 with the Gln/Arg mutation at position 24), showed solubility *in vivo* and *in vitro* and no formation of amyloid fibrils (12). Three 2-$\mu$s runs of SQNGNQQRG carried out with the same temperature and simulation protocols used for GNNQQNY did not show the formation of any stable in-register aggregates. Only three times did the SQNGNQQRG replicas aggregate into short-lived, parallel out-of-register $\beta$-sheets. These simulation results indicates that the force field and solvation model are not biased toward the formation of ordered aggregates. It is known that amino acid substitutions with charged residues can prevent fibril formation (28). The Arg at the C terminus might even prevent the early steps of in-register aggregation of SQNGNQQRG. On the other hand, Balbirnie *et al.* (12) have reported the *in vitro* formation of unbranched fibrils for the charged nonapeptide GNNQQNYQR. To investigate other possible reasons for the lack of ordered aggregates of SQNGNQQRG, an analysis of the out-of-register aggregates was performed. It was found that the $\phi$-dihedrals of the central glycines fluctuate on average 74°, which is much more than the average of 22° for the other nonterminal residues. These torsional fluctuations lead to the disruption of backbone hydrogen bonds and, finally, disaggregation. The high torsional mobility of the central glycine might also contribute to the prevention of full in-register aggregation even when the N-terminal residues are correctly adjusted. To further investigate whether the lack of aggregation events in the SQNGNQQRG runs is a consequence of the central glycine or charge repulsion, three additional 1-$\mu$s simulations were performed with the mutant peptide SENGNQQRG. Although the replacement of the first Gln by a Glu in SENGNQQRG should favor the antiparallel $\beta$-sheet by the Glu–Arg side-chain interactions, no in-register aggregates were observed (Table 1). Hence, the simulation results indicate that the flexibility of Gly disfavors the formation of ordered aggregates.

**Energy Surfaces.** For a system in thermodynamic equilibrium, the difference in free energy in going from a state A to a state B is proportional to the natural logarithm of the quotient of the probability of finding the system in state A divided by the probability of state B. The sampling of several transitions between disordered and ordered aggregates indicates that the simulations are close to equilibrium. Moreover, the free-energy surfaces constructed from two independent sets of 10 simulations have the same shape and show a low average error (see caption of Fig. 2). The free-energy surface has three distinctive minima at IP3, IM3, and the disordered aggregates, which in this projection includes the soluble state, i.e., conformations with one or all isolated replicas (Fig. 2 *Upper*). IP3 ($Q_p \geq 0.75$) is more stable than IM3 and has a free-energy difference of 2.8 kcal/mol from the disordered aggregates ($Q_p \leq 0.2$ and $Q_a \leq 0.2$), whereas the one between IM3 ($Q_p \geq 0.4$ and $Q_a \geq 0.4$) and the disordered aggregates is 3.3 kcal/mol. IP3 is also more stable than the out-of-register aggregate OM3, which collocates with IP2 ($Q_p \approx 0.5$ and $Q_a \leq 0.2$) on the energy surfaces. The free-energy difference of OM3 from the disordered aggregates was calculated from its population probability and is 3.1 kcal/mol.

The average effective energy $\langle E \rangle$ (sum of intrapeptide, interpeptide, and solvation energy) as a function of the fraction of in-register parallel, $Q_p$, and in-register antiparallel contacts, $Q_a$, shows an overall downhill landscape (Fig. 2 *Lower*). The three minima on this surface correspond to IP3, IM3, and IA3, with a most pronounced minimum for IP3. The effective energy does not include the configurational entropy of the peptide, which consists of conformational and vibrational entropy contributions (29). Hence, the free-energy minimum of the disordered aggregates originates from an entropic advantage because the effective energy is very unfavorable. On the first view, the barriers between the disordered and ordered aggregates also seem to have only an entropical origin because the average effective energy appears rather smooth. However, a closer look reveals two barriers on the effective energy surface, one at $Q_p = 0.7$ and $Q_a \leq 0.2$ and the other at $Q_p \leq 0.2$ and $Q_a = 0.7$. The preceding regions ($Q_p \approx 0.5$, $Q_a \leq 0.2$ and $Q_p \leq 0.2$, $Q_a \approx 0.5$) correspond to IP2 and IA2, respectively. As described above, IP2 and IA2 appear either alone or with a third strand that is not in register. The latter conformations have a lower effective energy than the former because of the supplementary interactions with the chain out of register. However, these interactions have to be broken to reach an aggregate with all three strands perfectly aligned. Indeed, this was observed in 21 of the 25 aggregation events of IP3. In 19 of the 21, two strands first aggregated in register followed by the third one, which was displaced by two or more residues. After 0.7 ns, on average, the out-of-register strand detached and rearranged in register. In the two remaining cases, the out-of-register strand was displaced by only one residue, i.e., OM3 was formed before the in-register aggregation. Hence, an enthalpic barrier has to be crossed when an assembly with out-of-register contacts converts to a fully in-register aggregate. These results are in contrast to a purely downhill surface observed for 20-residue peptides that fold to a three-stranded antiparallel $\beta$-sheet (16, 17). The absence of a connection between the three peptide replicas results in a larger number of different low-energy states and an effective energy surface that is more rough than for a three-stranded $\beta$-sheet peptide.

Backbone and side-chain interactions contribute to the enthalpic barriers as seen in Fig. 3. However, the contribution of the backbone interactions is more pronounced, because most out-of-register strands mainly form backbone contacts. It is interesting to note that the backbone hydrogen bonds slightly favor the antiparallel arrangement (Fig. 3 *Left*), in agreement with previous energy-minimization studies (30). On the other hand, the interactions between side chains clearly favor the parallel aggregate by about $-6.5$ kcal/mol (Fig. 3 *Right*). Recently, it has been proposed that the $\pi$ stacking between aromatic residues might contribute significantly to the thermodynamic stability of amyloid structures (31). In agreement with

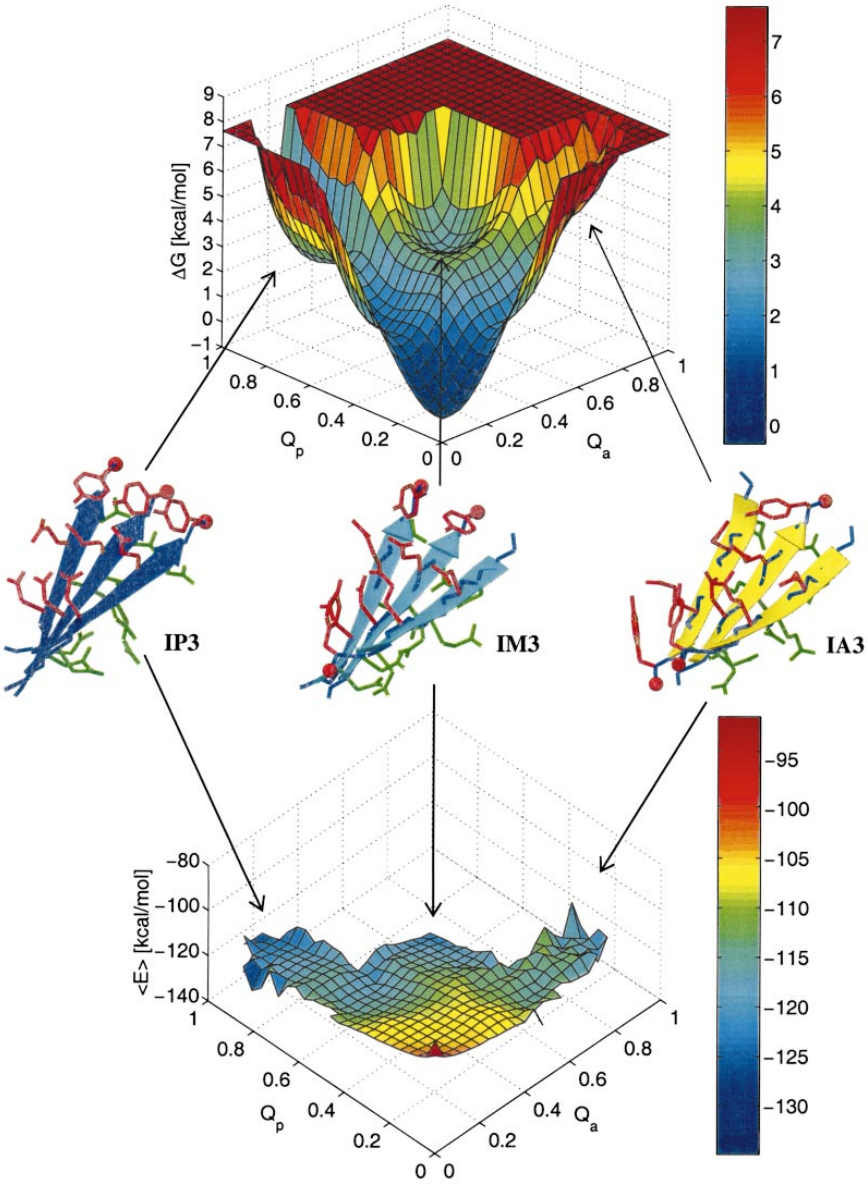


**Fig. 2.** Free energy ($\Delta G$, *Upper*) and average effective energy ($\langle E \rangle$, *Lower*) surface at 330 K as a function of the fraction of in-register parallel, $Q_p$, and in-register antiparallel, $Q_a$, contacts. A total of $10^6$ conformations sampled during the 20 simulations at 330 K were used. $\langle E \rangle$ was evaluated by averaging the effective energy values of the conformations within a bin without minimizing them. $\Delta G$ was computed as $-k_B T \ln(N_{n,m}/N_{0,0})$, where $N_{n,m}$ denotes the number of conformations with $n$ parallel and $m$ antiparallel contacts. The error in $\Delta G$ is estimated by separating the 20 simulations into two sets of 10 simulations each. The average error of $\langle E \rangle$ is 1.2 kcal/mol, and the average error of $\Delta G$ is 0.2 kcal/mol. Representative conformations of IP3, IM3, and IA3 sampled along the MD trajectories are shown in blue, cyan, and yellow, respectively.

this suggestion, the most favorable energetic contribution to the stability of IP3 originates from the Tyr–Tyr interactions (Table 2). The interaction energies of the tyrosines are significantly lower than the minimal stacking energy of −6.6 kcal/mol, recently calculated for an optimal Tyr–Tyr alignment (32). However, in the calculation of the minimal stacking energy the tyrosines were modeled by *p*-cresol, whose ring centroids could adopt a distance of 3.7 Å in the optimal stacking configuration. In the simulations at 330 K, by contrast, the average distance between the ring centroids is 5.2 Å in IP3. Nonetheless, the stacking interactions between the aromatic rings of tyrosine, as well as the higher number of hydrogen bonds between the side chains of polar residues, result in a clear preference of the parallel over the antiparallel aggregation of GNNQQNY. This

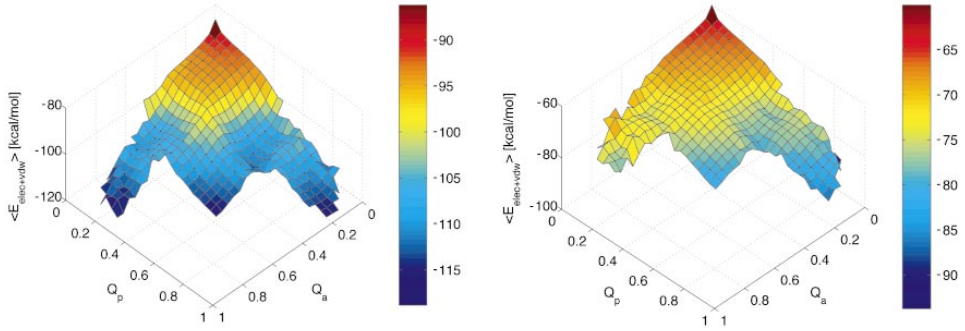


**Fig. 3.** Sum of van der Waals and electrostatic energies for the atoms in the backbone (*Left*) and in the side chains (*Right*) as a function of the fraction of in-register parallel, $Q_p$, and in-register antiparallel, $Q_a$, contacts. For clarity reasons, this plot has been rotated by 180° around a vertical axis with respect to Fig. 2.

behavior seems strongly related to the asymmetry in the sequence. One may therefore wonder whether a palindromic sequence prefers to form antiparallel aggregates. To test this hypothesis, three 1-$\mu$s simulations were carried out for the palindromic sequence GNNQQNNG. The three replicas of this peptide aggregated four times in the antiparallel and only once in the parallel arrangement (Table 1). These results indicate that backbone hydrogen bonds seem to turn the balance in favor of an antiparallel in-register aggregate if side-chain interactions are equally favorable in IP3 and IA3, as is the case for palindromic sequences.

**Influence of Tyrosine on the Early Aggregation Events.** Alanine substitution experiments on short fragments of the islet amyloid polypeptide (IAPP) and amyloid $\beta$ peptide ($A\beta$) showed that the aromatic residue Phe is crucial for their aggregation propensity (28, 33). It was proposed that interactions between aromatic residues might not only make a strong contribution to the thermodynamic stability of the amyloid structures but also provide order and directionality in the self-assembly. This hypothesis is investigated here by MD simulations of three mutants of the GNNQQNY peptide, which have no tyrosine (Table 1). If there is directionality in the self-assembly process, the peptides lacking the aromatic residue are expected to form in-register aggregates less frequently. The normalized frequencies to form IP2 (see *Materials and Methods*) are 18.5 ± 5.5 per $\mu$s for the wild-type peptide and 23.3 ± 8.9 per $\mu$s, 21.6 ± 4.4 per $\mu$s, and 15.9 ± 2.9 per $\mu$s for the GNNQQNA, GNNQQNG, and GNNQQN-mutant, respectively. The similarity in the frequencies to form IP2 aggregates indicates that the aromatic residue tyrosine does not provide more order to the aggregation process. On the contrary, GNNQQNA and GNNQQNG form IP2 aggregates slightly more often. However, the IP2 assemblies of the wild-type sequence are kinetically more stable than those of the mutants. The IP2 aggregate of the GNNQQNY peptide is stable for, on average, 3.2 ns, whereas the IP2 aggregate of the mutants disassociates already after 1.4 ns. The slower disaggregation of the wild-type peptide allowed that 9.7% of the 257 IP2 aggregates were elongated into IP3 by docking the third stand in register. This occurred for only 5.7% of the 226 IP2 aggregates formed in the mutant simulations.

**Table 2. Side-chain interaction energies in the IP3 conformation of GNNQQNY**

| Interacting pairs | $E_{elec}$* | $E_{vdW}$† | $E_{tot}$‡ |
|---|---|---|---|
| Tyr–Tyr | 0.0 | −2.3 | −2.3 |
| Asn–Asn | −1.3 | −0.8 | −2.1 |
| Gln–Gln | −0.5 | −1.0 | −1.5 |
| Gln–Tyr | −0.5 | −0.6 | −1.1 |
| Gln–Asn | −0.4 | −0.5 | −0.9 |

All energies are in kcal/mol.
*Electrostatic contribution to the interaction energy.
†van der Waals contribution to the interaction energy.
‡Total interaction energy.

Overall, the simulation results indicate that an aromatic residue does not give directionality to the self-assembly process but stabilizes the parallel aggregates. This increased stability gives the free strand more time to assemble in register. The aggregation process can formally be represented by the addition reaction (34)

$$S_1 + A_{n-1} \underset{\alpha_n}{\overset{\beta_n}{\rightleftharpoons}} A_n, \qquad [2]$$

where $S_1$ is an isolated peptide, $A_n$ ($n > 1$) is the aggregate containing $n$ peptides, and $\alpha_n$ and $\beta_n$ are the disaggregation and aggregation rate constants, respectively. The critical nucleus of aggregation is reached when the aggregation and disaggregation rates are the same. In the simulations presented here, the initial aggregation rate constants, $\beta_2$, are similar for the peptides with and without an aromatic residue. By contrast, the disaggregation constants, $\alpha_2$, are lower for the former. Although no predictions can be made for late rate constants ($n > 3$), the results suggest that, for a given monomer concentration, the peptides with a tyrosine reach the nucleus faster than the sequences lacking a tyrosine. Overall, the findings are consistent with the experimentally observed key role of aromatic residues in amyloid formation of peptides.

## Conclusions

The present study shows that it is possible to simulate with an atomic model the early steps of aggregation of an amyloid-forming peptide. The simulations give insights into the energetics of the early assemblies and the strong sequence dependence of aggregation. Backbone hydrogen bonds favor the antiparallel $\beta$-sheet packing but side-chain hydrogen bonds and aromatic stacking stabilize the in-register parallel aggregate. Simulations with peptides lacking the tyrosine indicate that aromatic residues lower the disaggregation rate of parallel assemblies. The dependence of aggregation and disaggregation rates on the sequence might be an essential factor determining the time scale on which

peptides reach the critical aggregation nucleus. Investigating other elements that influence nucleation is of major interest. Further studies of the dependence of aggregation kinetics on aromatic residues, charged residues (35), and peptide length by MD simulations might help to understand the nucleation of amyloidogenic peptides.

We are grateful to Prof. A. Baici and Dr. G. Settanni for helpful discussions and an anonymous referee for suggesting to run the control simulations with the mutant peptides GNNQQNA, GNNQQNNG, and SENGNQQRG. This work was supported by the Théodore Ott Foundation, the Swiss National Competence Center in Structural Biology, and the Swiss National Science Foundation (Grant 31-64968.01 to A.C.). J.G. is a fellow of the Swiss M.D.-Ph.D. program (Grant 3236-057617).

1. Dobson, C. M. (1999) *Trends Biochem. Sci.* **24,** 329–332.
2. Perutz, M. F. (1999) *Trends Biochem. Sci.* **24,** 58–63.
3. Blake, C. & Serpell, L. (1996) *Structure (London)* **4,** 989–998.
4. Malinchik, S. B., Inouye, H., Szumowski, K. E. & Kirschner, D. A. (1998) *Biophys. J.* **74,** 537–545.
5. Broglia, R. A., Tiana, G., Pasquali, S., Roman, H. E. & Vigezzi, E. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 12930–12933.
6. Giugliarelli, G., Micheletti, C., Banavar, J. R. & Maritan, A. (2000) *J. Chem. Phys.* **113,** 5072–5077.
7. Harrison, P. M., Chan, H. S., Prusiner, S. B. & Cohen, F. E. (1999) *J. Mol. Biol.* **286,** 593–606.
8. Smith, A. V. & Hall, C. K. (2001) *J. Mol. Biol.* **312,** 187–202.
9. Fernandez, A. & Boland, M. (2002) *FEBS Lett.* **529,** 298–302.
10. Ma, B. & Nussinov, R. (2002) *Protein Sci.* **11,** 2335–2350.
11. Ma, B. & Nussinov, R. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 14126–14131.
12. Balbirnie, M., Grothe, R. & Eisenberg, D. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 2375–2380.
13. Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J., Taddei, N., Ramponi, G., Dobson, C. M. & Stefani, M. (2002) *Nature* **416,** 507–511.
14. Ferrara, P., Apostolakis, J. & Caflisch, A. (2002) *Proteins Struct. Funct. Genet.* **46,** 24–33.
15. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983) *J. Comput. Chem.* **4,** 187–217.
16. Ferrara, P. & Caflisch, A. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10780–10785.
17. Ferrara, P. & Caflisch, A. (2001) *J. Mol. Biol.* **306,** 837–850.
18. Hiltpold, A., Ferrara, P., Gsponer, J. & Caflisch, A. (2000) *J. Phys. Chem. B* **104,** 10080–10086.
19. Gsponer, J. & Caflisch, A. (2001) *J. Mol. Biol.* **309,** 285–298.
20. Gsponer, J. & Caflisch, A. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6719–6724.
21. Ferrara, P., Apostolakis, J. & Caflisch, A. (2000) *J. Phys. Chem. B* **104,** 5000–5010.
22. Eaton, W. A., Munoz, V., Hagen, S. J., Jas, G. S., Lapidus, L. J., Henry, E. R. & Hofrichter, J. (2000) *Annu. Rev. Biophys. Biomol. Struct.* **29,** 327–359.
23. Cavalli, A., Ferrara, P. & Caflisch, A. (2002) *Proteins Struct. Funct. Genet.* **47,** 305–314.
24. De Alba, E., Santoro, J., Rico, M. & Jiménez, M. A. (1999) *Protein Sci.* **8,** 854–865.
25. Beglov, D. & Roux, B. (1994) *J. Chem. Phys.* **100,** 9050–9063.
26. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. (1977) *J. Comp. Phys.* **23,** 327–341.
27. Yang, A. S. & Honig, B. (1995) *J. Mol. Biol.* **252,** 366–376.
28. Tjernberg, L. O., Naslund, J., Lindqvist, F., Johansson, J., Karlstrom, A. R., Thyberg, J., Terenius, L. & Nordstedt, C. (1996) *J. Biol. Chem.* **271,** 8545–8548.
29. Lazaridis, T. & Karplus, M. (1999) *Proteins Struct. Funct. Genet.* **35,** 133–152.
30. Vasquez, M., Nemethy, G. & Scheraga, H. (1994) *Chem. Rev.* **94,** 2183–2239.
31. Gazit, E. (2002) *FASEB J.* **16,** 77–83.
32. Chelli, R., Gervasio, F. L., Procacci, P. & Schettino, V. (2002) *J. Am. Chem. Soc.* **124,** 6133–6143.
33. Azriel, R. & Gazit, E. (2001) *J. Biol. Chem.* **276,** 34156–34161.
34. Ferrone, F. (1999) *Methods Enzymol.* **309,** 256–274.
35. Lopez de la Paz, M., Goldie, K., Zurdo, J., Lacroix, E., Dobson, C. M., Hoenger, A. & Serrano, L. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 16052–16057.

# Part III

# Conclusion and Final Notes

# Chapter 8

# Conclusion

The present thesis focuses on two aspects. The first one is the development and implementation of implicit solvation models. The second one is their application to peptide folding and aggregation.

Implicit solvation models try to incorporate all solvent effects into a mean solvation energy term in order to eliminate the solvent degrees of freedom and to increase computational efficiency. Both atomic solvation energies and the screening have to be calculated in an implicit solvation model. A common choice to include the screening effect of the solvent is the use of a distance dependent dielectric function. The advantage of such a simple recipe is the computational efficiency. The drawback is that it can not discriminate between atoms that are buried inside the protein and feel the presence of solvation only marginally, and atoms that are near or on the surface and are strongly influenced by the solvent. The AEI model introduces a screening function that discriminates between buried and exposed atoms but stays close in efficiency to the distance dependent dielectric function. It utilizes only quantities that are simple to calculate. A measure for the amount of volume occupied by solute atoms around a pair of interaction centers is used to quantify the screening effect. It is demonstrated that exact (finite difference Poisson) interaction energies are well reproduced. The AEI model addresses only the interaction energy between charges and misses the self energies. The FACTS model is based on similar ideas and steric concepts as the AEI model but includes all electrostatic contributions to the solvation free energy. For each atom of the low dielectric solute the volume and the symmetry of the distribution of its neighbors are utilized to define a measure for how deeply an atom is buried inside the protein. This measure facilitates a fast and fully analytical calculation of atomic solvation energies which in turn can be used to obtain the screening by utilizing the Generalized Born formula. The FACTS model is thus a Generalized Born variant. Given its accuracy which is similar to the

most accurate Generalized Born implementations, the FACTS model represents by far the fastest Generalized Born implementation currently available. Both the AEI and FACTS models are developed from scratch, parameterized, and implemented in CHARMM as part of this thesis. They are only about three to four times slower than in vacuo molecular dynamics simulations.

The SASA model utilizes a distance dependent dielectric screening function, takes advantage of neutralized ionic side chains, and assumes that atomic solvation energies are proportional to the solvent accessible surface areas. It has been implemented in the official release of CHARMM as part of this thesis and is applied to peptide folding and aggregation. In the first study a three-stranded antiparallel $\beta$-sheet peptide is simulated for a total simulation time of 12.6 $\mu$s at the melting temperature of 330 K. During this time a total of 72 folding events are observed. It is demonstrated that the unfolded state ensemble contains many more conformers than those sampled during a single folding event. This confirms previous findings in lattice models that fast folding corresponds to a downhill process on a funnel-like free energy surface. The second study addresses aggregation properties of a small peptide. Experimental results suggest that all ordered aggregation processes have common key elements. Therefore, research on small and simplified systems that are able to form polymers in sheet configuration may provide valuable insight at atomic level into the pathologies related to protein deposits. The role of side-chain interactions in the early steps of the amyloid fibril forming process are investigated in this study. Aggregation MD simulations of prion-like peptides of the yeast protein Sup35 are performed. In agreement with experimental data they correctly generate the in-register parallel packing of $\beta$-strands. Energetically, backbone interactions favor the in-register antiparallel packing whereas hydrogen bond interactions between side-chains and stacking of aromatic residues favor the in-register parallel assembly. Overall the in-register parallel packing is more favorable and determines the statistically dominant configuration. The roughness of the effective and free energy surfaces is shown and it is demonstrated that the preferred pathway to the parallel aggregate does not correspond to a purely downhill profile on the effective energy surface. Additionally, the simulations confirm a strong sequence dependence of the aggregation kinetics.

# Chapter 9

# Final Notes

## 9.1 Acknowledgment

First and foremost I would like to thank very much Amedeo Caflisch for offering me the opportunity to work in the fascinating field of computational biology and for his continuous support throughout the Ph.D. in all aspects. It was a great pleasure to work with him. I appreciated a lot the atmosphere in the group and the excellent technical equipment. I am especially thankful for the freedom he gave me not only for choosing the research projects but also to take care of the computer cluster and to teach courses of mathematics.

Very special thanks I would also like to give Jörg Gsponer for personal and scientific support during the time we shared. I enjoyed very much working with him. A lot of thanks also go to Nicolas Majeux for a fruitful and comfortable collaboration and to Nicolas Budin for his scientific and personal help and hints. Many thanks go to Shaheen Ahmed who always took the time to help me when I came to him. Special thanks also go to Christian Bolliger for many scientific and non-scientific discussions and for some very cheerful events. Furthermore, many thanks for very stimulating scientific - and sometimes non-scientific - discussions go to Manuel Walker, Enrico Guarnera, Riccardo Pellarin, Philipp Werner, and Peter Kolb. Finally, I would like to thank all members of the group, former or present, in no particular order: Emanuele Paci, Joannis Apostolakis, Philippe Ferrara, Serena Fioravanti, Bernd Lüneburg, Francesco Rao, Gian Gaetano Tartaglia, Marco Cecchini, Giovanni Settanni, Andrea Cavalli, Michele Seeber, Raffaele Curcio, Gianluca Interlandi, Egon Perathoner, Marco Scarsi, Rainer Bökmann, and Catherine Tenette-Souaille.

I would like to give very warm thanks to my parents and my sister for their help in the most serious moments.

Finally, I would like to give many special thanks to my best friends Richard

Brogle, Olivier Sarbach, and Luciano Carraro (in no particular order), who I know from high school and who are always there when I need them.

## 9.2 Curriculum Vitae

### Personal

| | |
|---|---|
| Name | Urs Edgar Haberthür |
| Date of birth | 19 March 1972 |
| Place of birth | Winterthur (Switzerland) |
| Nationality | Swiss |

### Education & Research

| | |
|---|---|
| 1999-2004 | Ph.D. student in computational biophysics in the group of Prof. A. Caflisch at the University of Zurich. |
| 1992-1998 | Studies of physics at the Swiss Federal Institute of Technology (ETH) Zurich. Specialization in optimization algorithms, particle physics, and general relativity. Degree: Dipl. Phys. ETH. |
| 1987-1991 | Gymnasium (High School) in Winterthur (Switzerland) type C (specialization in natural sciences and mathematics). Degree: Matura. Honor: Graduation as best of type C. |
| 1985-1987 | Gymnasium (High School) in Winterthur (Switzerland). |

### Work Experience

| | |
|---|---|
| 2001-2004 | UNIX/Linux and Beowulf cluster system administrator for the *Computational Biomolecular Sciences Center* at the University of Zurich. Design, installation, configuration and maintenance of a 58 node Beowulf cluster in the research group of Prof. A. Caflisch. Development and implementation of a fully automatic installation and configuration environment. |
| 1999-2002 | Teacher of mathematics at the Institute for Operations Research at the University of Zurich. |
| 2000-2001 | Teacher for the practical training *Macromolecular Recognition: Computational Course in Biochemistry* (four terms). |
| 1996-1997 | Teacher of mathematics and for the practical training *Fluid dynamics* at the ETH Zurich (one term each). |
| 1992 | Employee at EMPA, the Swiss Federal Laboratories for Materials Testing and Research. |

# Part IV

# Appendix

# Appendix A

# Hints for Building a Beowulf Cluster

## A.1 Introduction

The workhorses of simulation science are computers. We currently have two Beowulf clusters. The first one, called santi cluster, consists of 33 rack mounted nodes that are divided into 32 slave nodes and 1 master node. The slave nodes are for number crunching and the master node manages the resources of the slave nodes. The master node is attached to a RAID 5 SCSI disk array with a capacity of about 500GB. (If one of the disk drives in a RAID 5 array goes down, the data can be recovered from the remaining drives.) Each slave node is a dual Athlon machine with 1.5GHz, 512MB of RAM, and a 40GB IDE hard drive. The master node is also a dual Athlon machine with 1.5GHz but features 1GB of RAM. The hard- and software of this cluster was fully set up by Transtec. Our overall experience with Transtec is good.

The second cluster, called mario cluster, consists of 56 desktop tower nodes (not rack mounted, also called boxes) that are simply put on shelves. This cluster is self-made. More precisely, we only bought the boxes and installed the rest on our own. Of the 56 machines, 42 are dual Athlon and 14 dual Pentium computers. The latest 32 nodes we bought in 2002 for the mario cluster are dual Athlon machines with 1.7GHz, 512MB of RAM, and a 60GB IDE hard drive each. The oldest nodes are 1GHz Pentium III computers. The master node is a dual Athlon machine with 1.4GHz, 1GB of RAM, and two 80GB IDE drives set up as a RAID 1 array. (RAID 1 means that the disks are mirrors of each other. No RAID 5 array is possible if only two drives are available.) This cluster also had a SCSI file server. We got rid of it since it was of no use any more (see below).

The santi cluster is homogeneous since all of it was bought at the same time in 2001 whereas the mario cluster is heterogeneous since it was installed and upgraded

over several years. What follows are hints that may be helpful to take decisions for building a Beowulf cluster. Most of our experience we got from building the mario cluster.

## A.2 Location

### A.2.1 Cooling

Most important is the cooling of the room where the cluster stays. The ideal room temperature is between 20°C and 22°C. The cooling capacity needed depends on the number and type of machines that are installed. Athlon machines produce considerably more heat than Pentium machines. We measured that our most demanding dual Athlon nodes consume about 270VA each if fully loaded, roughly 80% is burnt as heat. We have a 12.5kVA(in)/10.5kW(out) air conditioner for the complete mario cluster. It works fine but the air conditioner is close to its limit. As to our experience, companies don't emphasize the cooling problem enough from their side. The general assumption is that you have enough cooling capacity for what you order.

### A.2.2 UPS

Since there are always power failures, it is recommended to put all machines on a UPS right from the start. It is much more complicated to connect machines to a UPS at a later time since one has to shut them down, rearrange cables, etc. The choice of the appropriate UPS(s) is a matter of its own. In principle, there are two possibilities: one large UPS or several small ones. The large ones (10kVA upwards) are huge and heavy, measuring something like 1.5x1.5x1.5m and weighting several hundreds of kilos. These large UPSs are designed to cover an interrupt of 10 minutes and more so they need many batteries. The small ones (about 3kVA) measure, for instance, 20x44x55cm and weight about 60kg. One can connect about 10 dual Athlon machines on one of them and it can cover a break of about 1 minute. The small ones are of course much more comfortable to handle and also much cheaper, even if you have to buy several ones to serve the same number of nodes as the large one does. The choice of the UPS(s) depends on how the power supply is organized at the institute. In our case, if we have a power failure, after about 30s the emergency generator of the university supplies the power. So we chose the small UPSs.

### A.2.3 Space

For installing and maintaining the cluster it is comfortable if the room is not too small. There is always some work to be done in the cluster room, not only for installation. From time to time, a node has to be taken out to be repaired (hard disk failure, broken network card, and the like). To our experience, it is a good idea to be generous not only with the space but also with the length of all the cables (network cables, power cables etc.). The basic rule seems to be: One always needs more than one thinks.

### A.2.4 Network

The room should feature fast network connections (possibly more than 10MBit) and preferably more than just one. Although one may be enough in theory, in practice more are needed, for instance to access online help in the cluster room via a notebook.

## A.3 Hardware

The basic question is whether or not one needs a shared memory machine. Shared memory machines are much more expensive that non-shared memory ones. To our knowledge non-shared memory machines are fine for molecular dynamics (MD) simulations. CHARMM only uses little memory (about 1% on a 512MB of RAM machine). If one does Poisson calculations or similar things, one needs more memory depending on the grid spacing, size of the molecule, etc. Up to now we were always able to do all our calculations with 512MB of RAM per node. If one decides to install a non-shared memory machine, a Beowulf cluster is the state of the art choice: It is fast and cheap. A Beowulf cluster typically consists of one master node, many number crunching (slave) nodes, and a file server.

### A.3.1 Rack Mounted Cluster versus Cluster of Boxes

If one sets up a Beowulf cluster, one has to decide whether one would like to install boxes or a rack mounted cluster. Roughly speaking, boxes are about half the price of a rack mounted system (according to the offers we got in summer 2001). Boxes also have the advantage that when after two or three years of operation one replaces them (or some of them), they can be recycled within the university. They can be given to secretaries, used for less computation expensive tasks that experimentalists may need in the laboratory, etc. One can't recycle rack mounted

computers so easily - they are too noisy for office use, for instance. However, rack mounted clusters require less space and look cleaner. But by now there are also cupboards for boxes available that make it look good. See, for instance, helics.iwr.uni-heidelberg.de/services/equipment/helics/gallery/index.php. For our next upgrade phase of the santi cluster we plan to no longer buy rack mounted nodes but also boxes.

### A.3.2 Test Machine

It can happen that on some particular board and chip set combinations, CHARMM crashes randomly with no motivation. Therefore, before buying a complete cluster from a specific company, it is crucial to buy only one machine first that is exactly identical in hardware to the other machines that are planned to be delivered. Then one has to test CHARMM on this single machine. We usually fill the test machine with at least two CHARMM jobs for about one week. If CHARMM doesn't crash during this period, the hardware is probably fine. In our group we have come to the conclusion that CHARMM is actually the best test for the hardware: If CHARMM runs fine, any software runs fine. We once didn't do such a test and then had to have replaced 14 motherboards because they weren't compatible with CHARMM. Although the company had to pay for it, we of course had all the trouble of taking out the machines from the cluster room, etc.

### A.3.3 Offers

One always should ask for offers from several companies. By now we have machines from three different companies for the mario cluster. Two are local to Switzerland (www.0-8-15.ch and www.dalco.ch) and the other one is Transtec that operates in England too. Our overall experience with Transtec is good. In comparing the different offers one has to pay close attention to the service included. Anything but a three year on site guarantee is not good enough in our opinion. Note that there will always be hardware failures (especially in the first weeks, but also later) and if one doesn't have on site guarantee, one will have to send or carry the machines somewhere which is extremely impractical. This point must not be underestimated. International companies are for instance Dr. Koch (www.koch-computer.de), Transtec (www.transtec.com), and Alineos (www.alineos.com).

### A.3.4 Computers

#### A.3.4.1 Slave Nodes

Our slave nodes typically contain the motherboard, two Athlon CPUs (the fastest ones with 1.7GHz), 512MB of RAM, a standard 100MBit network adapter, a hard drive, floppy drive, and in the case of the mario cluster a graphics card. None of them features a CD or DVD drive. Usually the slave nodes of a Beowulf cluster don't include a graphics card (in particular if the cluster is rack mounted) since the output can be displayed via a console server, a special hardware device. In this case the serial interfaces of the slave nodes are connected to the console server which itself is connected to the master node. Using telnet one can connect to a specific port of the console server (every port of the console server belongs to a slave node) and then the monitor of the master node works as if it was connected directly to the slave node. However, in this way only text can be displayed and the node has to be connected to the console server. All our boxes do include a graphics card and we don't have a console server for the mario cluster. We found that it is very convenient to be able to remove a node from the cluster room and to have a closer look at it outside to do some diagnosis or other work on it. Also, if one plans to recycle the boxes later on for other use, a graphics card is necessary. We always simply chose the cheapest graphics card. For the mario cluster we bought a small flat screen that can easily be moved around in the cluster room. If a node has a problem and we don't want to take it out from the cluster, we simply connect the monitor and keyboard to that node.

#### A.3.4.2 Master Node

The master nodes of the mario and santi cluster have similar hardware like the slave nodes but 1GB of RAM, more hard disk space, and a DVD drive. The DVD drive is important for the software installation.

#### A.3.4.3 File Server and Backups

The idea of a file server is to store the data of the people centrally. It is usually a machine with a SCSI array, i.e., an array of something like 10 SCSI disks set up as RAID 5. RAID 5 means that the data are stored redundantly: If one disk fails, no data is lost. But SCSI disks are very expensive, up to four times the price of IDE disks. Manufacturers claim that SCSI disks are more reliable than IDE disks and designed for 24h operation whereas IDE disks are not. However, fact is that both SCSI and IDE disks do fail and it is hard to judge from daily life whether or

not SCSI disks are more reliable. Some say SCSI is more reliable, some say IDE is more reliable. As to our knowledge and experience, reliability does not favor the one or the other. There are arguments for SCSI, and these are access time and rate of data transfer. These are important if one hosts a web server or data bank, but in our opinion not to store MD data. If one would like to buy a file server, a good choice may be to go for an IDE array with an IDE2SCSI adapter that mimics a SCSI interface. There are external IDE2SCSI solutions (Arena, Infortrend) and internal ones (Adaptec, 3Ware, AMI) available. The companies will make appropriate offers if asked. If data are stored centrally on a file server, the data are easily organized and backups can easily be made. Part of the SCSI array of the santi cluster is backed up automatically by a backup robot. This service is offered by the university and can only be used by machines in certain locations. The mario cluster can't use this service.

However tempting the idea of a file server may look in general, we have come to the conclusion that it is not appropriate for our needs. Actually, these days we decided to get rid of our mario file server for two reasons. The first problem is that even if you insert 20 disks in the file server, it will not be enough. It seems that disks always get filled, no matter how many there are. The second problem is that if everybody copied the trajectory files from the MD simulations from the nodes to the file server and then did the analysis on the file server, the machine would be massively overloaded.

Therefore, we adopted a different procedure for the mario cluster: Every slave node has at least a 60GB IDE disk. These add up to terabytes of storage capacity. The users of the cluster are assigned these disks as storage disks where they can put their data. Furthermore, trajectory files from simulations we usually analyze locally on the node where they were produced and don't copy them around. Each user is responsible for backups of his own. Important data can be copied on several disks, burnt to CDs, or written to tape.

### A.3.5 Size of the Cluster and Processor Speed

Two or three years after having bought some machines, they are no longer competitive. Therefore we adopted the scheme that we only buy few machines at a certain moment in time but we keep short intervals between upgrades. Thus our latest hardware is always state of the art. Once or even twice a year we buy new machines. This year we bought 32 nodes, next year we will probably buy another 32. Other groups buy a cluster of 500 nodes all at the same time but then they lack the money for upgrades every year. Also, we never bought the fastest available

processor but rather the second fastest one because the difference in price never justified the gain in speed. The scale is not linear.

### A.3.6 Compatibility with the Operating System

If one plans to install a free operating system like Linux one has to assure that the hardware is supported. Usually the companies check this if asked, but generally it is a good idea not to buy the latest network adapter, for instance, but a slightly older one for which the drivers have been already in use for some time.

### A.3.7 Network and Switch

The way the nodes are connected which each other can strongly affect the performance of the cluster, in particular for parallel jobs. Explicit water simulations are usually run in parallel. Both our clusters feature Ethernet. We don't have Myrinet or any other more advanced network technology. We found after extensive tests (done by an external diploma student at a different institute) that Myrinet does not improve the performance of CHARMM for parallel jobs significantly, although the costs for the hardware and human time to install and maintain it rise considerably.

For the mario cluster we have a small 16 port switch and a 48 port CISCO switch that features its own operating system. It can be managed remotely. While this is certainly desirable for complicated networks, it is an overkill for a cluster like mario or santi. It even complicates the setting up since it needs to be configured correctly. The santi cluster features two 3COM switches. Our experience is that Ethernet is fine for MD with CHARMM, also for parallel jobs with explicit water, and that it's suitable to go for good switches, avoiding both the low and high end area. Before installing an advanced network technology a careful evaluation of the needs, benefits, and costs is necessary.

## A.4 Software of the Mario Cluster

### A.4.1 Automated Installation

On the mario cluster we run SuSE Linux 7.3. The master node is set up such that installation and configuration of any slave node is completely automatic. Adding a new slave node includes the following steps: The slave node is powered on with a boot disk in the floppy drive. This boot disk contains only a driver for the network card and a small program that sends a query for a DHCP server. The master node

answers the query by giving the slave node its IP address, hostname, IP address of a name server, IP address of a gateway, and the name of a file that the slave node has to download from the master node. The master node identifies the slave node uniquely via its MAC address that it also broadcasted in the query of the slave node. (The MAC address first needs to be added in the DHCP configuration file of the DHCP server on the master node.) Then the slave node downloads the file from the master node which is a bootable Linux kernel. This is the standard etherboot procedure. With this kernel, the slave node starts the installation program (autoyast1) that is mainly downloading RPM packages from the master node and installing them. After having finished installing the packages, autoyast1 executes user defined scripts that update the node and configure it for the use in the cluster. The information needed to install and configure a specific node is organized centrally in sets on the master node so as to give us the largest flexibility for handling a heterogenous system. Each node is assigned to certain sets. A set in this installation scheme is similar to a class in object oriented programming. A set encapsulates data (configuration files, for instance) and actions (scripts, for instance) that have to be performed on that data (for instance, copy the configuration files from the master node to a specific location on the slave node). The largest set consists simply of all nodes. A subset consists, for instance, of all Pentium machines, another one of all Athlon machines, and again another one of all slave nodes, etc. Each set contains all the information that is only relevant for the machines it contains. Different hardware may require different partitioning, different boot kernels, different boot parameters, etc. Furthermore, what packages should be installed and what user defined scripts should be run on the nodes may also vary.

We tested several distributions and programs to set up an automated installation and found SuSE to be the most comfortable one. In particular, the SuSE DVD contains a huge amount of packages so one doesn't have to download them from the net, compile them, etc. Furthermore, autoyast1 works nicely. Since SuSE 8.0, autoyast1 has been replaced by autoyast2 which we haven't tested yet. It is XML based. Since our setup for the installation does the configuration with user defined scripts and autoyast1 is used only for the raw installation (partitioning, installing RPM packages, and some basic system configuration), we don't expect too many troubles when switching to autoyast2.

### A.4.2 Maintaining and Using the Cluster

Many configuration files like the /etc/passwd, /etc/shadow, or the /etc/hosts file are organized centrally. They are only on the master node and accessed by the slave

nodes via NIS (Network Information System). To submit jobs to the slave nodes we use OpenPBS, the non-commercial version of the PBS Batch Queuing System. It takes care of distributing the jobs on the slave nodes. bdsh (bash distributed shell) is a tool that allows for easy and fast execution of a single command or whole scripts on all nodes simultaneously by only starting it on the master node. This is very comfortable to find files, copy files etc. All mounting is done via NFS (Network File System), using automount that is much safer than statical NFS mounts. Automounting means that a filesystem is mounted only when it is accessed and if it has not been used for, say, 5 minutes, it is unmounted again. Contrary to the santi cluster and against common advice, any machine in the mario cluster can be mounted on any other machine in the cluster. The advantage is the ease with which files can be accessed. The disadvantage is that if a machine crashes, other machines may also hang. But this situation only rarely arose in our case so we consider the gain worth the price. Also contrary to the santi cluster, all the software that is installed on the master node is installed on each slave node too (and this is a substantial amount of software). Thus anything that can be done on the master node can also be done on the slave nodes. Users appreciate in particular that they can run any analyzing tool, any graphical tool (like xmgrace, etc.) simply anywhere. It saves the work and network traffic arising from copying files around or access them via NFS. To run parallel CHARMM jobs we use MPI (Message Passing Interface). MD simulations with explicit water, for instance, we only run in parallel, involving at least two processors.

## A.5 Software of the Santi Cluster

The santi cluster runs Debian Linux. The main difference to the installation scheme of the mario cluster is that each slave node only contains a very small Linux system of about 50MB that is stored only in a RAM disk and not on the hard drive. When a santi slave node is booted, it also sends a DHCP query and the file the slave node has to download from the master node is an image of this RAM disk that contains a mini Linux operating system. The advantage compared to the mario cluster is that installation is much faster (since there is nothing that really needs to be installed). The disadvantage is that the slave nodes can only be used for number crunching whereas each slave node of the mario cluster features about 3GB of software. The overall configuration of the santi cluster is similar to the mario cluster. The santi cluster also uses NIS, NFS with automount, OpenPBS, and dsh instead of bdsh.

# Appendix B

# The CHARMM Documentation of the SASA Implicit Solvation Model

## B.1 Characteristics of the SASA Model

The SASA model is a fast implicit solvation model that is useful to simulate structured peptides and miniprotein motifs [1]. The polar and non-polar contributions of each atom to the free energy of solvation are assumed to be proportional to their solvent accessible surface areas. The SASA model uses only two surface-tension like solvation parameters (constants of proportionality) and approximates the solvent accessible surface area of each solute atom with a simple analytical function that is easily derivable. The electrostatic screening between solute charges is accounted for by using a distance dependent dielectric function and by neutralizing the formal charges (Asp, Glu, Arg, Lys, and the termini) as in the EEF1 model [2].

The SASA model has been successfully applied to peptides, removing the major artifacts of in vacuo simulations and reproducing reversible folding [3]. Benchmarks indicate that a simulation with SASA is only about 50% slower than an in vacuo simulation.

## B.2 Range and Limitations

The SASA model has been applied to structured peptides, see for instance [3]. However, it should not be used for large proteins mainly for two reasons. Firstly, it has been parameterized for small proteins [1] and secondly, the dielectric function does not take different environments into account, i.e., it does not distinguish whether or not the interacting partial charges are buried or on the protein surface.

## B.3 Theoretical Aspects

The potential energy of the system consisting of the solute and the solvent can be decomposed in three parts: the intra-solute potential energy U(X), the intra-solvent potential energy V(Y), and the interaction potential energy of the solute and the solvent W(X,Y), where X denotes the degrees of freedom of the solute and Y the degrees of freedom of the solvent. Integrating out all solvent degrees of freedom one obtains the potential of mean force W(X), also called the effective energy. It can be written as the sum of the intra-solute potential energy U(X) and the free energy of solvation, or mean solvation term, DW(X) that describes all solvent induced effects. This is a rigorous result from statistical mechanics. For more details consult, for instance, the review [4].

Any implicit solvation model based on the solvent accessible surface area approximates the major contribution to the free energy of solvation using the first solvation layer, and the screening effect of the solvent. Following this idea it is assumed that the mean solvation term is a sum of atomic contributions proportional to the solvent accessible surface area of each atom, plus an energy term that accounts for the screening. The constants of proportionality are the surface-tension like solvation parameters.

## B.4 Technical Aspects of the SASA Model

There are three important aspects for any implicit solvation model that is based on the solvent accessible surface area: (A) how to calculate the atomic solvent accessible surface areas, (B) how many surface-tension like solvation parameters to use, and (C) how to account for the screening of solute charges.

### B.4.1 (A) Calculation of the Solvent Accessible Surface Area

In the SASA model the solvent accessible surface area is approximated by a probabilistic approach. For details please refer to [5]. The base is a simple formula for the probability to hit the accessible surface of atom $i$ if $N$ atoms are present by choosing randomly a spot on the solvation shell of atom $i$. However, this formula is only true under the assumption that all the atoms are distributed randomly. This is of course not the case because of covalent geometry and the Pauli exclusion principle. Now instead of elaborating in sophisticated probability calculations the pragmatic approach of [6] is adapted where the original formula is parameterized by including two sets of parameters: a set of probabilistic parameters (called atom type parameters in

[1,6]) and a set of connectivity parameters. The probabilistic parameters depend on the atom type and correct for systematic errors primarily due to hybridization. The connectivity parameters distinguish bound atoms from more distant ones. Also, the radii used to calculate the approximated solvent accessible surface were optimized for this purpose. (These radii are used only for the calculation of the approximated area and they are different from the radii used for the CHARMM van der Waals energy term.) The optimal values for the radii, probabilistic parameters, and connectivity parameters were taken from [6]. The atom types in [6] and the CHARMM atom types do not match exactly, so the most reasonable assignments were chosen, analogous to the choices in [7].

The only internal degree of freedom considered by this approximation is the interatomic distance. Therefore the major defect is the absence of a better correlation between exact and approximated areas upon changes in internal coordinates like dihedral angles.

### B.4.2 (B) Solvation Parameters

By default, the SASA model uses only two non-vanishing surface-tension like solvation parameters: one for hydrophobic and one for hydrophilic groups. The solvation of explicit hydrogen atoms is neglected. However, this can be changed by the user (see below). The solvation parameters were adjusted for the EEF1-modified CHARMM 19 polar hydrogen parameter set by Philippe Ferrara in a trial and error approach in 1999. The criterion was to minimize the root mean square deviation from the native state for six small proteins by performing molecular dynamics simulations of 1ns at 300K [1].

### B.4.3 (C) Screening of Solute Charges

For the screening of solute charges the SASA model uses a distant dependent screening function, $\varepsilon(r) = 2r$, and neutralizes the charged groups of polar amino acids in exactly the same way as it is done in the EEF1 model of Lazaridis and Karplus [2].

## B.5 Implementation in CHARMM

To every CHARMM atom type the SASA model assigns a surface-tension like solvation parameter (zero by default for explicit hydrogen atoms and non-zero for hydrophobic and hydrophilic groups), a radius optimized for the approximation of the solvent accessible surface area, and a probabilistic parameter. Additionally,

a connectivity parameter is assigned to every pair depending on whether the two atoms of the pair form a 1-2 or a more distant pair. These parameters are called the SASA parameters [1].

When initializing SASA, i.e., by invoking the SASA command, all the SASA parameters are printed to the CHARMM output file. A value of -999.000 means that this specific parameter hasn't yet been determined for SASA and therefore the corresponding CHARMM atom type can't be used by default for the SASA calculations - the user would have to assign a meaningful value. (Currently, SASA does not support the following CHARMM atom types by default: HA, HT, LP, CT, CM, NP, OH2, OM, OT, OS, and FE.) All SASA parameters can be changed by the user. For the corresponding syntax, see below.

To evaluate the approximative formula for the solvent accessible surface area of an atom $i$, all neighbors of atom $i$ within a certain cutoff are required. This cutoff is calculated by 2*(2.365Å+1.4Å)=7.53Å, where 2.365Å is the largest van der Waals radius in the CHARMM parameter set 19, and 1.4Å is the radius of the solvent probe sphere. Most, but not all of these neighbors, are included in the nonbond pair list in CHARMM. The missing pairs are stored in a new pair list, the SASA pair list. More precisely, the SASA pair list contains all pairs that (1) are not in the nonbond pair list, that (2) do not belong to the fixed exclusions (as given in the topology file), and that (3) are within the above mentioned cutoff. The SASA pair list assures correct operation of the SASA model for any nonbond exclusion mode (-5 to 5). This means that for any nonbond exclusion mode from 1 to 5, the SASA energy and its derivatives are identical. The same is true for any nonbond exclusion mode from -5 and 0. The differences in the SASA energy and its derivatives between the nonbond exclusion modes regions of -5 to 0 and 1 to 5 stem from the fact that the fixed exclusions are treated differently: For an exclusion mode from -5 to 0 they are included in the nonbond pair list, opposed to an exclusion mode from 1 to 5 where they are excluded from the nonbond pair list.

## B.6 Caveat

Please note that the radii used for the calculation of the approximated solvent accessible surface area are labeled 'van der Waals radii' in the SASA output of the CHARMM output file. This is consistent with the terminology used in [6]. However, these radii are different from and do not replace in any way the CHARMM default van der Waals radii used to calculate the van der Waals energy term.

## B.7 Additional Input Files

Two additional files are needed to use SASA (taken from EEF1):

1. toph19_eef1.inp: This is a modification of toph19.inp where ionic side chains and termini are neutralized and which contains an extra parameter type (CR).

2. param19_eef1.inp: This is a modification of param19.inp which includes the extra parameter type (CR).

These files can be found in test/data/.

## B.8 Syntax of the SASA Command

There is only one SASA command:

- SASA atom-selection [S<number> <real>] [R<number> <real>] [P<number> <real>] [fcon <real>] [ncon <real>] [surf] [infx]

- atom-selection:== (see *note select:(chmdoc/select.doc).)

- <number>:== number corresponding to a CHARMM atom type from param19.inp or param19_eef1.inp according to table B.1.

The SASA command sets up the SASA model for a simulation. All the values have to be given on one command line. Invoking the SASA command a second time reinitializes all values either to the default or to the user specified values.

- atom-selection: This determines the atoms to be used for the SASA calculations. All atoms that are not included in this selection are treated by SASA as if not existent. Their solvation free energies are not considered, i.e., are set to zero, and the decrease in the solvent accessible surface areas of the selected atoms due to the not selected ones is neglected.

- S<number> <real>: Changes the surface-tension like solvation parameter of the CHARMM atom type corresponding to <number> (see list above) from the default value to <real>.

- R<number> <real>: Changes the radius used by SASA (for the calculation of the approximated solvent accessible surface areas) of the CHARMM atom type corresponding to <number> (see table B.1) from the default value to <real>.

| Number | Atom type |
|---|---|
| 001 | H |
| 002 | HC |
| 003 | HA |
| 004 | HT |
| 005 | LP |
| 006 | CT |
| 007 | C |
| 008 | CH1E |
| 009 | CH2E |
| 010 | CH3E |
| 011 | CR1E |
| 012 | CM |
| 013 | N |
| 014 | NR |
| 015 | NP |
| 016 | NH1 |
| 017 | NH2 |
| 018 | NH3 |
| 019 | NC2 |
| 020 | O |
| 021 | OC |
| 022 | OH1 |
| 023 | OH2 |
| 024 | OM |
| 025 | OT |
| 026 | OS |
| 027 | S |
| 028 | SH1E |
| 029 | FE |
| 030 | CR |

Table B.1: Numbers and atom types

- P<number> <real>: Changes the probabilistic parameter of the CHARMM atom type corresponding to <number> (see table B.1) from the default value to <real>.

- fcon <real>: Changes the connectivity parameter for 1-2 pairs from the default value to <real>.

- ncon <real>: Changes the connectivity parameter for more distant than 1-2 pairs from the default value to <real>.

- surf: The approximated atomic solvent accessible surface areas are stored in WMAIN.

- infx: Includes the fixed exclusion pairs in the SASA pair list. By default, the fixed exclusion atoms are not considered in the SASA surface calculations for historical reasons. This means that, for instance, for the CG of the residue PHE (see the topology file), CZ is not considered to calculate its accessible surface. To include all neighbors use the keyword infx, but note that SASA was not parameterized with this option.

The SASA standard setup [1,3] looks like this:

```
nbond nbxmod 5 atom rdiel shift vatom vdistance vshift -
cutnb 8.0 ctofnb 7.5 ctonnb 6.5 eps 2.0 e14fac 0.4 wmin 1.5

sasa selection (.not.  hydrogen) end
```

The nonbond options are the default nonbond options from the default param19.inp file with the exception of the eps value that is set to 2 instead of 1 and rdiel is used instead of cdiel. The default solvation parameters are $-0.06$ for hydrophilic groups (N, NR, NH1, NH2, NH3, NC2, O, OC, and OH1) and 0.012 for hydrophobic groups (C, CH1E, CH2E, CH3E, CR1E, S, SH1E, and CR) and 0.0 for explicit hydrogen atoms. Please note that the param19_eef1.inp file has different default nonbond options (especially the cutoffs are different) that were not used to parameterize SASA and therefore should not be used or used only with care in simulations with SASA since consistency is lost.

If you select all atoms for SASA instead of excluding the (explicit) hydrogens, all atoms will be considered for calculating the solvent accessible surface area of each atom. However, since the solvation parameter for hydrogen atoms is zero by default, the solvation energy of the hydrogen atoms is also zero by default. If you insist on including the solvation energy due to the solute hydrogen atoms, you have

to assign a non-zero solvation parameter to the hydrogen atoms by yourself, using 'S001 <real> S002 <real>' in the CHARMM input file. Be aware that this is not the default.

## B.9 Solvation Parameters for Proteins (not Fully Tested)

A second set of surface-tension like solvation parameters has been derived: $-0.144$ for hydrophilic groups, 0.024 for hydrophobic groups and 0.0 for explicit hydrogen atoms with $\varepsilon(r) = r$. It is not the default. It seems to work better for large proteins but no results have been published up to date (June 2004).

## B.10 More than One Chain

If you run a simulation with several molecules (so that you have more than one segment identifier), make sure that you invoke the SASA command after the generation of the last segment since any molecule generated after the last use of the SASA command is not included in the SASA calculations. You don't have to invoke the SASA command after every generation of a segment, it is sufficient to use the SASA command once after the generation of the last segment.

## B.11 Accessing the Solvation Energy

The SASA solvation energy is stored in the variable 'SASL'. Use '?SASL' in the CHARMM input file to access the value.

## B.12 New Surface Parameters

A new set of surface parameters (radii, probabilistic parameters, and connectivity parameters) was derived in 2002 by Haberthuer. They give a better correlation with exact analytical surfaces than the original Hasel and Still surface parameters. (A set of 20 structures [10 native and 10 unfolded conformations] was used for the calibration.) The new surface parameters are not the default because the surface-tension like solvation parameters were optimized by Ferrara with the original Hasel and Still surface parameters. The following example illustrates how to setup SASA with the new surface parameters.

```
nbond nbxmod 5 atom rdiel shift vatom vdistance vshift -
cutnb 8.0 ctofnb 7.5 ctonnb 6.5 eps 2.0 e14fac 0.4 wmin 1.5

sasa infx selection (all) end -
r001 0.100001 r002 0.491498 r007 1.369460 r008 1.895900 -
r009 2.286480 r010 2.651430 r011 2.075670 r013 1.895900 -
r014 0.323677 r015 1.390860 r017 0.100001 r018 1.469600 -
r019 1.554210 r020 1.552220 r021 1.681850 r022 1.652130 -
r023 1.548960 r028 2.188920 r029 1.878290 r031 1.578670 -
p001 0.988957 p002 1.777570 p007 1.304200 p008 1.316430 -
p009 1.182780 p010 1.083020 p011 1.134370 p013 1.316430 -
p014 1.130220 p015 1.510700 p017 1.592510 p018 1.261620 -
p019 1.187100 p020 1.053870 p021 1.036120 p022 1.139990 -
p023 1.096650 p028 1.680130 p029 0.907302 p031 1.276250 -
fcon 0.342979 ncon 0.507482
```

## B.13 References


- [1] Ferrara, P.; Apostolakis, J.; Caflisch, A.; Evaluation of a Fast Implicit Solvent Model for Molecular Dynamics Simulations; Proteins 2002; 46; 24-33.
- [2] Lazaridis, T.; Karplus, M.; Effective Energy Function for Proteins in Solution; Proteins 1999; 288; 477-487.
- [3] Ferrara, P.; Caflisch, A.; Folding Simulations of a Three-Stranded Antiparallel Beta-Sheet Peptide; Proc. Natl. Acad. Sci. USA; 2000; 97; 10780.
- [4] Roux, B.; Simonson, T.; Implicit Solvent Models; Biophysical Chemistry; 78; 1999; 1-20.
- [5] Wodak, S. J.; Janin, J.; Analytical Approximation to the Solvent Accessible Surface Area of Proteins; Proc. Natl. Acad. Sci. USA; 1980; 77; 1736.
- [6] Hasel, W.; Hendrickson, T. F.; Clark, S. W.; A Rapid Approximation to the Solvent Accessible Surface Area of Atoms; Tetrahedron Computer Methodology 1988; Vol. 1; No. 2; 103-116.
- [7] Fraternali, F.; van Gunsteren, W. F.; An Efficient Mean Solvation Force Model for Use in Molecular Dynamics Simulations of Proteins in Aqueous Solution; J. Mol. Biol. 1996; 256; 939.

## B.14 Examples

Check test/c29test/sasa.inp for a more complex example with minimization, equilibration and dynamics. Here is a short version:

```
* Example input file for the SASA implicit solvation model.  *

!  --- Begin generation procedure ---
open read card name toph19_eef1.inp unit 30
read rtf card unit 30
close unit 30

open read card name param19_eef1.inp unit 30
read parameter card unit 30
close unit 30

open read card name filename.crd unit 30
read sequence coor unit 30
close unit 30

generate main warn setup
!  --- End generation procedure ---

!  --- Begin reading coordinates ---
open read card name filename.crd unit 30
read coordinate card unit 30
close unit 30
!  --- End reading coordinates ---

!  --- Begin setting up SASA ---
!  Use the SASA standard setup.
nbond nbxmod 5 atom rdiel shift vatom vdistance vshift -
cutnb 8.0 ctofnb 7.5 ctonnb 6.5 eps 2.0 e14fac 0.4 wmin 1.5

sasa selection (.not.  hydrogen) end
!  --- End setting up SASA ---

!  --- Begin minimization ---
minimize sd nstep 300 nprint 20 tolgrad 0.1
minimize conj nstep 200 nprint 20 tolgrad 0.1
!  --- End minimization ---

stop
```